Statistical Methods to Adjust for Measurement Error in Risk Prediction Models and Observational Studies

Statistical Methods to Adjust for Measurement Error in Risk Prediction Models and Observational Studies

A dissertation presented

by

Danielle Braun

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

December 2013

Dissertation Advisor: Professor Giovanni Parmigiani Danielle Braun

Statistical Methods to Adjust for Measurement Error in Risk Prediction Models and Observational Studies

Abstract

The first part of this dissertation focuses on methods to adjust for measurement error in risk prediction models. In chapter one, we propose a nonparametric adjustment for measurement error in time to event data. Measurement error in time to event data used as a predictor will lead to inaccurate predictions. This arises in the context of self-reported family history, a time to event covariate often measured with error, used in Mendelian risk prediction models. Using validation data, we propose a method to adjust for measurement error in this setting. We estimate the measurement error process using a nonparametric smoothed Kaplan-Meier estimator, and use Monte Carlo integration to implement the adjustment. We apply our method to simulated data in the context of Mendelian risk prediction models and multivariate survival prediction models, and illustrate our method using a data application for Mendelian risk prediction models. Results show our adjusted method corrects for measurement error mainly in two aspects; by improving calibration and total accuracy. In some scenarios discrimination is also improved. In chapter two, we use the methods proposed in chapter one to extend Mendelian risk prediction models to handle misreported family history.

The second part of this dissertation focuses on methods to adjust for measurement error in observational studies. In chapter three, we propose various methods to adjust for a mismeasured exposure using validation data and propensity scores. Propensity score methods assume that the treatment assignment is error-free, but in reality these variables can be subject to measurement error. This arises in the context of comparative effectiveness research, in which accurate procedural codes are not always available. When using propensity score based methods, this error affects the treatment assignment variable di-

rectly, as well as the propensity score. We propose a two step maximum likelihood approach using validation data to adjust for measurement error. In addition, we show the bias introduced when using error-prone treatment in the inverse probability weighting estimator and propose an approach to eliminate this bias. Simulations show our proposed approaches reduce bias and mean squared error of the treatment effect estimator compared to using the error-prone treatment assignment.

Contents

Acknowledgments

I would like to thank my thesis advisor, Giovanni Parmigiani. I am very fortunate to have worked with him over the last several years. He taught me both how to approach and solve statistical problems, and about the importance of the clinical relevance of a problem and collaborative research. Giovanni gave me the freedom to pursue areas of research that I found interesting, which I very much appreciate. I am grateful for all of his patience and support.

I would also like to thank Malka Gorfine for all of her guidance, help, and encouragement. She was extremely generous with her time and I greatly appreciate all of her feedback and insight. In addition, I would like to thank my other committee members, Tianxi Cai and Donna Spiegelman, for their helpful questions and suggestions.

I am fortunate to have had the opportunity to work with a number of great collaborators over the past few years. The first part of my thesis work was done in collaboration with Hormuzd Katki and Argyrios Ziogas. The second part of my thesis work was done in collaboration with Corwin Zigler and Francesca Dominici.

I would also like to thank the BayesMendel lab group: Amanda Blackford, Su-Chun Cheng, Jonathan Chipman, Emanuele Mazzola, and Lorenzo Trippa. This group provided me with great feedback and comments over the years.

Finally, I want to thank my mom and sister for their support.

Nonparametric Adjustment for Measurement Error in Time to Event Data

Danielle Braun, Malka Gorfine, Hormuzd Katki, Argyrios Ziogas, and Giovanni Parmigiani

Department of Biostatistics

Harvard School of Public Health

1.1 Introduction

Measurement error in binary and continuous covariates has been studied extensively in the literature (Carroll, 2006a, among others). The focus of this chapter is on measurement error in time to event data which are used as covariates in a model. Time to event data are coded by two variables; T indicating either time to event or censoring whichever occurs first, and δ indicating whether the event occurred. We focus on scenarios in which both T and δ are measured with error. Because of the relationship between T and δ, standard techniques adjusting for measurement error in binary or continuous covariates cannot be applied directly.

Previous work in this setting has focused on measurement error in survival outcomes (Meier et al., 2003). Meier et al. consider a discrete setting in which subjects are tested at predetermined time points until the time of first observed failure. Using the sensitivity and specificity rates of failing, they develop a model for the measurement error process based on a validation data set, and incorporate this into an adjusted proportional hazards model. Their method cannot be extended to our setting for two reasons. The first, that our time to event data is not obtained by repeated testing, instead we look at scenarios for which the time to event data is measured with error at one time point. The second, that our interest is in time to event data used as covariates in the model and not as outcomes. We are not aware of any literature directly applicable to our setting.

We focus on scenarios for which a prediction model based on error-free time to event data has been developed, however, implementation of these prediction models uses error-prone time to event data. We came across this problem in Mendelian risk prediction models, which use Mendelian laws of inheritance to calculate the probability that an individual carries a cancer causing inherited mutation based on family history, known mutation prevalence, and penetrance (the probability of having a disease at a certain age given the mutation status) (Murphy and Mutalik, 1969). These models assume family history is error-free, however, in practice they often rely on self-reported family history, which is not always accurate. The accuracy of self-reported family history has been evaluated in several studies which show that sensitivity and specificity estimates for reported disease

status vary by degree of relative and type of cancer. For example, for breast cancer in first-degree relatives sensitivity estimates vary from 65% to 95% while specificity estimates are usually around $98\% - 99\%$ (Mai et al., 2011; Ziogas and Anton-Culver, 2003). The effects of misreported family history on Mendelian risk prediction models have been examined by Katki (2006). Both errors in underreporting of disease status and rounding of age were considered, and it was shown that misreporting of family history, especially in disease status, leads to distortions in predictions. A model based on these inaccurate predictions will not be well calibrated.

Although our work is motivated by the setting of Mendelian risk prediction models, it is applicable to other scenarios, particularly in the context of multivariate survival prediction models. For example, suppose one has developed a model predicting overall survival (OS) based on error-free progression free survival (PFS). In reality however, PFS is often error-prone due to two main reasons; assessment of tumor size based on radiological scans varies by the observer and scans are taken at regularly scheduled intervals (Korn et al., 2010). Gray et al. (2009) evaluated measurement error in the PFS endpoint by comparing PFS assessment by an independent review facility (IRF) (which would be considered the error-free PFS) to an investigator-based assessment (which would be considered the error-prone PFS). They conducted an independent review of trial E2100, an open-label multi-center, randomized, phase III trial conducted by the Eastern Cooperative Oncology Group (ECOG). They saw that for 6% of the patients a PFS event was only identified by IRF, and for 18.1% of the patients a PFS event was only identified by local review. 43.5% of patients had PFS events identified by both IRF and local review, and for those the date of PFS was the same for 54.5% of the patients and within 6 weeks for 70.4% of the patients.

In this context, suppose one has developed a model to predict OS based on IRF determined PFS. In practice, one might want to use the prediction model to predict OS using PFS determined by local review as a covariate and not by IRF, since local review might be the only feasible option. Another example could arise in the context of using short term survival as a predictor for long term survival (Parast et al., 2012). If one has developed a model based on error-free short term survival outcomes, but in practice these are

measured with error, our proposed method is again applicable.

In section 1.2 we formulate our proposed method. No assumptions are being used for the measurement error model, thus we consider our proposed method as non-parametric. We apply our proposed approach to Mendelian risk prediction models in section 1.3, and to other multivariate survival prediction models in section 1.4. Simulation results are presented in section 1.5. Finally, we illustrate our method using a data application in the context of Mendelian risk prediction models, in section 1.6, and summarize the main results in section 1.7.

1.2 Model

1.2.1 Notations

We consider a setting in which we have an outcome Y and time to event data which are used as covariates in the model. More specifically, let T^o be the true failure time, let C be the true right-censoring time, $T = min(T^o, C)$ and $\delta = 1(T^o \leq C)$. We denote the error-free predictor as $H = (T, \delta)$. We assume a model $P(Y|H)$ has been developed elsewhere based on this true time to event data.

Now, suppose when implementing this model the time to event data used as covariates has error. We denote this error-prone predictor as $H^* = (T^*, \delta^*)$. We also assume we have a validation study which includes both the error-free time to event data, H, and the error-prone time to event data, H^*. We do not assume the validation data includes Y.

1.2.2 Proposed Method

Our goal is to predict the outcome Y based on the observed data H^*, namely to estimate $P(Y|H^*)$. We propose to rewrite $P(Y|H^*)$ by applying the total law of probability and Bayes rule, as follows:

$$P(Y|H^*) = \int_H P(Y, H|H^*)\, dH = \int_H P(Y|H, H^*)P(H|H^*)\, dH = \int_H P(Y|H)P(H|H^*)\, dH \tag{1.1}$$

The measurement error process, $P(H|H^*)$, is modeled in the validation study. Depending on the data, integrating over all possible values of H might be computationally challenging. In these cases, we propose using Monte Carlo integration, using $P(H|H^*)$ to select samples.

We propose to treat $H^* = (T^*, \delta^*)$ as covariates, and model the measurement error process using a survival distribution assuming conditional independence of event and censoring times given T^*, δ^*:

$$P(T, \delta|T^*, \delta^*) = \lambda(T|T^*, \delta^*)^{\delta} S(T|T^*, \delta^*) h(T|T^*, \delta^*)^{1-\delta} G(T|T^*, \delta^*). \tag{1.2}$$

λ and S are the hazard and survival functions of the event time, and h and G are the hazard and survival functions of the censoring time. Based on the validation data, one could estimate the survival distribution $P(T, \delta|T^*, \delta^*)$ parametrically (for example using a Weibull distribution or accelerated failure time model), semi-parametrically (for example using a Cox model), or non-parametrically (for example using Kaplan-Meier estimators). In general, we recommend using smoothed Kaplan-Meier estimators (Beran, 1981), requiring no parametric assumptions on the data. Specifically, we stratify by δ^*, and estimate each conditional survival function while borrowing information from the neighboring observations based on the values of T^*.

1.2.3 Model Assumptions

The last equality in Equation (1.1) follows from the surrogacy assumption. We assume H^* is a surrogate for H; in other words H^* contains no information on predicting Y in addition to the information already contained in H. This is plausible in our setting since the probability of the outcome conditional on both the error-free and error-prone predictor should only be influenced by the error-free predictor.

Our proposed approach assumes that the measurement error model $P(H|H^*)$ is transportable; that the measurement error model observed in the validation study can be applied to the population of interest. For this to be true, the validation and the target population should be as similar as possible. One should give thought to the choice of an appropriate validation study when applying our proposed method.

1.3 Mendelian Risk Prediction Models

Mendelian risk prediction models have been previously developed, and details of these models can be found elsewhere (Berry et al., 1997; Parmigiani et al., 1998). Briefly, these models calculate an individual's carrier probability $P(\gamma_0|H)$, where $H_i = (T_i, \delta_i)$, $H = (H_0, H_1, ..., H_R)$ for a family of size R, and the consultand is denoted as 0. $\gamma_i = (\gamma_{i1}, ..., \gamma_{iM})$, where $\gamma_{im} = 1$ indicates carrying the genetic variants that confer disease risk for each individual i at a gene $m = 1, .., M$, and $\gamma_{im} = 0$ otherwise. Furthermore, let $T_i = min(T_i^o, C_i)$ where T_i^o is the age of disease diagnosis for individual i, C_i is the current age or age of death for individual i, and $\delta_i = \mathbf{1}(T_i^o \leq C_i)$ for individual i.

Using Bayes rule and assuming conditional independence of family members' phenotype given their genotypes, we can write the consultand's carrier probability as follows;

$$P(\gamma_0|H_0, H_1, .., H_R) = \frac{P(\gamma_0) \sum_{\gamma_1,...,\gamma_R} \prod_{i=1}^{R} P(H_i|\gamma_i) P(\gamma_1, .., \gamma_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0) \sum_{\gamma_1,...,\gamma_R} \prod_{i=1}^{R} P(H_i|\gamma_i) P(\gamma_1, .., \gamma_R|\gamma_0)}. \tag{1.3}$$

These models are typically developed using validated family history, H. That is, penetrance estimates for these models are based on a meta-analysis. We expect the majority of the studies in the meta-analysis use validated family history. Software for performing these calculations is available as part of the BayesMendel R package (Chen et al., 2004), which includes BRCAPRO, a model identifying individuals at high risk of breast and ovarian cancer, MMRPro a model identifying individuals at high risk of Lynch Syndrome, and PancPRO a model identifying individuals at high risk of pancreatic cancer.

In practice, instead of having the true history, H, a consultand has a reported history H^*. Our goal is to estimate $P(\gamma_0|H^*)$. Using Equation (1.1), this probability can be rewritten as follows:

$$P(\gamma_0|H^*) = \int_H P(\gamma_0|H) P(H|H^*)\, dH. \tag{1.4}$$

$P(\gamma_0|H)$ is then calculated using Equation (1.3). The measurement error process, $P(H|H^*)$, is estimated by using smoothed Kaplan-Meier estimators based on a validation study, and the integration is implemented using a Monte Carlo integration.

In Mendelian risk prediction models, our interest is not only in estimating carrier prob-

abilities, but also in estimating the probability of the consultand surviving until time t, $P(T^o > t|H^*)$. A Mendelian model for $P(T^o > t|H)$ has been previously developed, and is also available as part of the BayesMendel R package. Similarly, using Equation (1.1), our proposed method can be implemented in this context: $P(T^o > t|H^*) = \int_H P(T^o > t|H)P(H|H^*)$ dH.

1.4 Multivariate Survival Prediction Models

The problem of measurement error in time to event data can also be applicable to the context of other multivariate survival prediction models. We continue our discussion of a hypothetical example in the context of predicting OS. Assuming a prediction model for OS using error-free PFS has been developed, one might be interested in implementing these models using error-prone PFS (since error-prone PFS might be the only information available). In this scenario, we let T_1^o denote OS time, $H = (T_2, \delta_2)$ be error-free PFS, and $H^* = (T_2^*, \delta_2^*)$ be error-prone PFS. Suppose one has developed a prediction model, $P(T_1^o > t|H)$, based on a study using error-free PFS, our interest however is in predicting $P(T_1^o > t|H^*)$. Using Equation (1.1) we can rewrite this probability as: $P(T_1^o > t|H^*) = \int_H P(T_1^o > t|H)P(H|H^*)\, dH$. $P(H|H^*)$ would be estimated using validation data containing both error-free PFS and error-prone PFS. Studies such as the one conducted by Gray et al. (2009), would be a good source of validation data, since they compared PFS assessment conducted by IRF review (error-free PFS) to PFS assessment conducted by local review (error-prone PFS). Similarly, this notation can also be applicable if we were interested in predicting long term survival time (T_1^o using our notation), based on an error-prone short term survival (H^* using our notation), assuming a model for $P(T_1^o > t|H)$ has been developed elsewhere.

1.5 Simulations

1.5.1 Mendelian Risk Prediction Models

In the context of Mendelian risk prediction models, for each simulation scenario two data sets were simulated; the first is used to model the measurement error process (the validation study), and the second to which we apply our method and estimate the carrier probability of each consultand given their family history. For the validation data set, $100,000$ families with 5 members (mother, father, and three daughters) were simulated. These simulations focus on one gene, BRCA1, and only on breast cancer. Carrier probabilities, $P(\gamma = 1)$, were assumed to be 0.006098, which is the allele frequency for BRCA1 in the Ashkenazi Jewish population. Error-free breast cancer failure times were simulated for each member based on known penetrance. We used the same penetrance estimates used by BRCAPRO, which are based on a meta-analysis. Error-free censoring times were simulated from a normal distribution with mean 55 and standard deviation 10. A large validation data is used because the allele frequency for BRCA1 is low. Since breast cancer failure times were generated using the penetrance, a large data set is needed in order to have a decent number of breast cancer events in the data.

Two settings for the measurement error in disease status were considered. The first using sensitivity=0.954 and specificity=0.974, taken from Ziogas and Anton-Culver (2003), the second using sensitivity=0.649 and specificity=0.990, taken from Mai et al. (2011). Four settings for the measurement error in age were considered. For the first three settings, we assume an additive classical model; $T^* = T + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, and $\sigma = 5$, $\sigma = 3$, and $\sigma = 1$. For the fourth setting, we assume a multiplicative measurement error model, $T^* = TU$, where $U \sim \exp(1)$. $P(H|H^*)$ was estimated using smoothed Kaplan-Meier with the nearest neighborhoods kernel using the prodlim R package (Gerds, 2011). The optimal bandwidths were calculated using the direct plug in approach proposed by Sheather and Jones (1991).

We applied our method to $50,000$ consultands for which their family history was generated in a similar manner. Specifically, for each of the $50,000$ families, carrier probability for the consultand based on family history was calculated using BRCAPRO for three dif-

ferent approaches; based on the simulated error-free history $P(\gamma_0|H)$, based on the simulated error-prone history by naively replacing H by H^*, denoted by $\tilde{P}(\gamma_0|H^*)$, and using our adjusted model $P(\gamma_0|H^*) = \int_H P(\gamma_0|H)P(H|H^*)\,dH$ (Table 1.1). For our adjusted model approach, Monte Carlo integration was applied by sampling 100 configurations of H based on $P(H|H^*)$.

These approaches were evaluated using three different measures; ratio of observed to expected events (O/E), mean squared error of prediction (MSEP), and area under the response operating characteristics curve (ROC-AUC). Model calibration is evaluated using the ratio of observed to expected events, which can be written as: $\frac{\sum_{i=1}^{n}\mathbf{1}(\gamma_i=1)}{\sum_{i=1}^{n}\hat{P}_i}$, where $\hat{P}_i = \widehat{P_i(\gamma_0|H)}$ for the O/E based on the error-free family history, $\hat{P}_i = \widehat{\tilde{P}_i(\gamma_0|H^*)}$ for the O/E based on the error prone family history, and $\hat{P}_i = \widehat{P_i(\gamma_0|H^*)}$ for the O/E based on the adjustment approach. Overall accuracy is evaluated using MSEP, which is the mean of the squared differences between the $\hat{P}_i$ and the error-free predictions. Therefore, MSEP based on the error-free data is always 0. For the error-prone predictions MSEP is calculated by $\frac{1}{n}\sum_{i=1}^{n}(\widehat{\tilde{P}_i(\gamma_0|H^*)} - \widehat{P_i(\gamma_0|H)})^2$, and for the adjustment approach it is calculated by $\frac{1}{n}\sum_{i=1}^{n}(\widehat{P_i(\gamma_0|H^*)} - \widehat{P_i(\gamma_0|H)})^2$. Model discrimination is evaluated using ROC-AUC. We used the verification R package (Gilleland, 2009) which calculates ROC-AUC following the process outlined by Mason and Graham (2002). ROC-AUC was calculated using the three different predictions (error-free, error-prone, and based on the adjustment approach).

The ratio of observed to expected events (O/E) based on the error-free family history is close to 1 in all simulation settings. The O/E based on the error-prone family history is lower than one, when using sensitivity=0.954 and specificity=0.974, and higher than one when using sensitivity=0.649 and specificity=0.990 (with the exception of the scenarios involving multiplicative error in age, for which the O/E is always less than one). When the sensitivity is lower, we have more underreporting of disease, which drives the O/E to be greater than one since the expected probabilities are lower. In all simulations, adjusting the error-prone improves the O/E substantially, and shifts it so it is closer to 1. For example, in the first simulation scenario, the O/E based on error-free family history

is 0.9773, based on error-prone family history it is 0.8190, and based on the adjustment it is 0.9712. Thus, we are able to eliminate almost all the bias.

MSEP based on adjusting the error-prone data is lower than MSEP based on the error-prone data alone for all simulations, by an amount that varies but can be substantial (Table 1.1). For example, in the first simulation scenario the square root of the MSEP multiplied by 1000, based on the error-prone family history is 19.1351, and based on the adjustment is 16.7405. ROC-AUC in all simulations are higher based on the error-free data compared to the error-prone data. In simulations involving additive error in age, ROC-AUC values are about the same using our adjustment compared to the error-prone data alone. For example, in the first simulation scenario ROC-AUC was 0.8160 based on error-free family history, 0.8090 based on error-prone family history, and 0.8086 based on the adjustment. In simulations involving a multiplicative error in age, where the error in age is stronger, ROC-AUC is higher using our adjustment compared to the error-prone data alone. For example, in the fourth simulation scenario, ROC-AUC was 0.8145 based on error-free family history, 0.7185 based on error-prone family history, and 0.8020 based on the adjustment.

Plots of predictions based on error-free, error-prone, and the adjustment are shown for two scenarios. The first, shown in red, in the top panel of Figure 1.1, corresponds to the first row in Table 1.1, representing a simulation setting with sensitivity 0.954 and specificity 0.974. The second, shown in blue, in the bottom panel of Figure 1.1, corresponds to the fifth row in Table 1.1, representing a simulation setting with sensitivity 0.649 and specificity 0.990. The first column on the left, shows predictions based on error-free family history compared to error-prone family history. There is more underreporting of cancer (in the plot these are individuals who are below the 45^o line) in the second scenario, due to the lower sensitivity, and more over-reporting (in the plot these are individuals who are above the 45^o line) in the first scenario, due to the lower specificity. In both scenarios, we have many individuals close to the 45^o line, corresponding to simulated families for which disease status was equivalent using error-prone and error-free family history. The second column in the middle, shows predictions based on error-free family history compared to our adjustment. In both scenarios, we see more individuals below the 45^o line, implying

Table 1.1: Mendelian Risk Prediction Simulation Results. MSEP and O/E improve using the adjusted proposed method, ROC-AUC either improves or remains the same depending on the setting.

consultand	Sens/Spec	Error in Age	$\sqrt{MSEP^{\dagger}} * 1000$			O/E			ROC-AUC		
			Error-Free	Error-Prone	Adjusted	Error-Free	Error-Prone	Adjusted	Error-Free	Error-Prone	Adjusted
Mother	0.954, 0.974	a: $N(0, 5^2)$	0.0000	19.1351	16.7405	0.9773	0.8190	0.9712	0.8160	0.8090	0.8086
		a: $N(0, 3^2)$	0.0000	18.1006	15.9490	0.9773	0.8280	0.9746	0.8160	0.8098	0.8078
		a: $N(0, 1^2)$	0.0000	17.5526	15.6430	0.9773	0.8327	0.9746	0.8160	0.8102	0.8115
		m: $exp(1)$	0.0000	43.2855	21.3037	0.9833	0.6063	0.9484	0.8145	0.7185	0.8020
	0.649, 0.990	a: $N(0, 5^2)$	0.0000	21.3122	20.8859	0.9773	1.0466	0.9783	0.8160	0.7814	0.7803
		a: $N(0, 3^2)$	0.0000	20.7947	20.5099	0.9773	1.0556	0.9792	0.8160	0.7821	0.7815
		a: $N(0, 1^2)$	0.0000	20.5213	20.1459	0.9773	1.0604	0.9737	0.8160	0.7826	0.7818
		m: $exp(1)$	0.0000	36.0314	23.8184	0.9817	0.8026	0.9565	0.8155	0.7140	0.7752
Daughter	0.954, 0.974	a: $N(0, 5^2)$	0.0000	18.8437	16.5614	0.9719	0.8166	0.9680	0.8171	0.8070	0.8082
		a: $N(0, 3^2)$	0.0000	17.7033	15.6918	0.9719	0.8256	0.9659	0.8171	0.8083	0.8086
		a: $N(0, 1^2)$	0.0000	17.1421	15.1775	0.9719	0.8301	0.9680	0.8171	0.8093	0.8083
		m: $exp(1)$	0.0000	43.4717	21.0365	0.9785	0.6028	0.9445	0.8162	0.7146	0.7976
	0.649, 0.990	a: $N(0, 5^2)$	0.0000	20.1613	20.0763	0.9719	1.0573	0.9872	0.8171	0.7895	0.7862
		a: $N(0, 3^2)$	0.0000	19.6759	19.5498	0.9719	1.0661	0.9834	0.8171	0.7913	0.7904
		a: $N(0, 1^2)$	0.0000	19.4507	19.1871	0.9719	1.0708	0.9803	0.8171	0.7928	0.7911
		m: $exp(1)$	0.0000	35.2381	23.0851	0.9760	0.8087	0.9535	0.8165	0.7229	0.7745

† MSEP: difference between (adjusted)error-prone and error-free predictions.

a: indicates a classical additive model; $T^* = T + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, m: indicates a multiplicative measurement error model, $T^* = TU$, $U \sim exp(\lambda)$.

our adjustment method slightly over adjusts by shifting probabilities down. The third column on the right, shows predictions based on error-prone family history compared to our adjustment. We can see that especially in the first scenario (which has more over-reporting of cancer than the second scenario), our adjustment shifts individuals' carrier probabilities down.

Overall, the adjustment method improves MSEP and calibration in all scenarios, while in some scenarios ROC-AUC remains the same (in the scenarios of additive error in age), in other scenarios the adjustment method improves ROC-AUC (in the scenarios of multiplicative error in age scenarios).

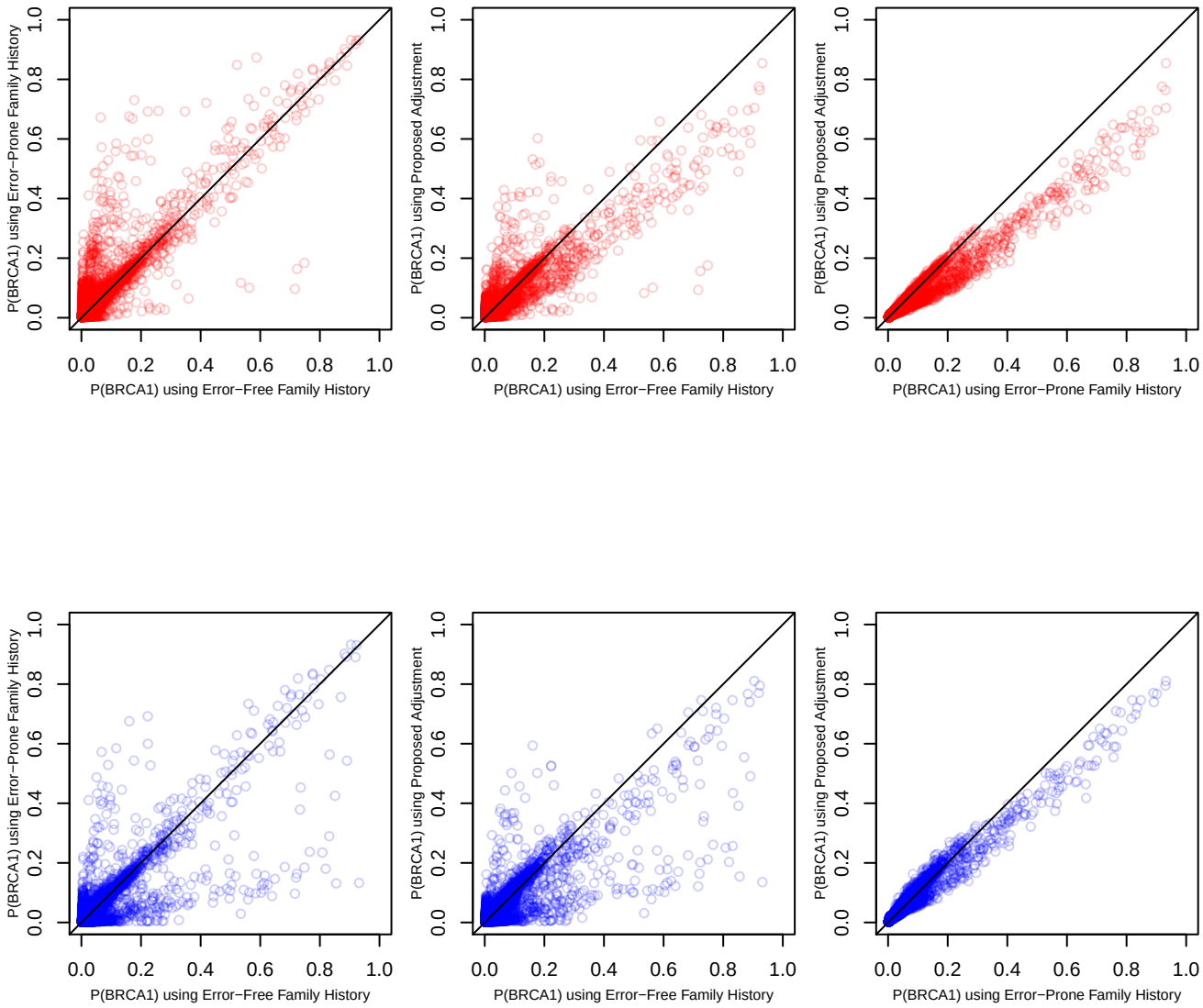


Figure 1.1: P(BRCA1) for simulated families based on error-free family history, error-prone family history, and proposed adjustment. In red, a simulation setting with sensitivity 0.954 and specificity 0.974. In purple, a simulation setting with sensitivity 0.649 and specificity 0.990.

1.5.2 Multivariate Survival Prediction Models

Simulations were also preformed in the context of predicting OS based on PFS. For these simulations, we consider a hypothetical scenario in which one has developed a prediction model for OS using error-free PFS as a predictor. In reality, however, PFS is measured with error, as shown in the E2100 review conducted by Gray et al. (2009). Based on the results of this review, Korn et al. (2010), performed simulations to assess the potential bias of measurement error in PFS on the conclusions of a proportional hazards analysis of a randomized trials. We follow a similar approach in generating the data for our simulations. Three data sets were generated for each simulation scenario. The first data set was generated to obtain a prediction model for OS based on error-free PFS. We assume a scenario in which patients were followed for progression free survival for a period of 25 months. Some of these patients will have a progression event during this time interval, whereas others will be censored. We assume that censoring represents the patient's last visit to the clinician's office. The goal is to predict the overall survival of a patient, t months from the time they either have a progression event or are censored.

For this first data set, we generate OS as well as error-free PFS for $1,000$ patients. For these simulations $1,000$ patients provide a large enough validation data, since the event rate is high (around 45%). We begin by simulating progression free survival times based on a Weibull distribution with shape parameter=1.456 and scale parameter=11.063 (based on E2100) . We assume a censoring distribution between 0 and 25 months with density $f(t) = 2(25 - t)/625$, resulting in approximately 55% censoring. To generate the overall survival times, for those who had a progression event, we assume the probability of death in this subpopulation is 50%. We generate a time for this death event based on an additive model; death time=progression time+$|(N(12, 5^2))|$. For those who did not have a progression event, we assume the probability of death in this subpopulation is 10%. We generate a time for this death event based on an additive model; death time=last clinician visit+$|(N(60, 5^2))|$. For those who did not have a death event, for simplicity we assume no censoring occurs, and assign them a survival time equal to their progression free survival time plus t months (where t is fixed and equal for everyone). We fit a prediction model

$P(T_1^o > t + T_2 | T_2, \delta_2)$ using smoothed Kaplan-Meier estimators for $\delta_2 = 0$ and 1 separately. The second data set was generated to model the measurement error process (the validation study). For this data set, we generate error-free PFS as well as error-prone PFS for $1,000$ individuals. We simulate error-free PFS as we did in the first data set. We then generate error-prone PFS from the error-free PFS. We introduce error in PFS events using various sensitivities and specificities (E2100 had 88% sensitivity and 64% specificity). We introduce error in PFS time as follows. If both error-free PFS and error-prone PFS are events: 55% of the time we assume agreement in the time of the event (based on E2100). For those who don't have complete agreement in the time of the event, 35% were within 6 weeks of each other. Therefore, we use multiplicative measurement error with log-normal distribution with standard deviation parameter $\log(1.5)$ (Korn et al., 2010). If both error-free PFS and error-prone PFS are non-events, we assume no error in PFS time. If error-free PFS is an event but error-prone PFS is not an event, we assign the error-prone PFS time to be the censoring time. If error-free PFS is not an event but error-prone PFS is an event, we assign the error-prone PFS time to be the minimum of the simulated progression failure time and end of study (25 months). We obtain error-free and error-prone PFS, and determine the measurement error model, $P(T_2, \delta_2 | T_2^*, \delta_2^*)$ using smoothed Kaplan-Meier estimators based on this simulated data set.

The third data set was generated to apply the proposed adjustment method to. For the third data set, we generate error-prone and error-free PFS as well as OS, for $1,000$ individuals as we did for the first two data sets. We preform prediction calculations on this data set. We are able to compare three different prediction calculations; the first using the error-free PFS as a covariate and calculating $P(T_1^o > t | T_2, \delta_2)$, the second by naively replacing the error-free PFS by the error-prone PFS as a covariate and calculating $\tilde{P}(T_1^o > t | T_2^*, \delta_2^*)$, and the third using our measurement error adjustment: $P(T_1^o > t | T_2^*, \delta_2^*) = \int_H P(T_1^o > t | T_2^*, \delta_2^*) P(T_2, \delta_2 | T_2^*, \delta_2^*) \, dH$. All three prediction calculations were done for a fixed $t = 60$.

In addition, we compare our proposed method to an alternative approach of modeling the measurement error process. This approach, assumes the measurement error is not dependent on time, in other words $P(T_2, \delta_2 | T_2^*, \delta_2^*) = P(\delta_2 | \delta_2^*)$. We estimate $P(\delta_2 | \delta_2^*)$ (NPV

and PPV) in the validation study, and use it to adjust for measurement error as follows; $P(T_1^o > t|T_2^*, \delta_2^*) = \int_H P(T_1^o > t|T_2, \delta_2)P(\delta_2|\delta_2^*)\, dH$. We will refer to this as the time independent adjustment.

Simulations for various values of sensitivity (varying in increments of 0.1 from 0.1 to 1) and specificity (varying in increments of 0.1 from 0.1 to 1) were conducted. The O/E ratios based on error-free covariates is close to one, and becomes even closer to one as sensitivity increases (Figure 1.2). The O/E ratios based on the error-prone covariates, decrease as sensitivity increases, and increase as specificity increases. A low sensitivity corresponds to underreporting of events, corresponding to higher O/E ratios, whereas a low specificity corresponds to over-reporting of events, corresponding to lower O/E ratios. Thus, as sensitivity increases O/E decreases, and as specificity increases O/E decreases. The right combination of sensitivity/specificity can lead to ad O/E close to 1. Based on the full adjustment method, O/E ratios are very close to one and do not vary much across sensitivity and specificity (Figure 1.2). O/E ratios based on the time independent adjustment increase slightly as sensitivity increases, and decrease slightly as specificity increases.

The full adjustment preformed best in terms of MSEP compared to both no adjustment and the time independent adjustment (Figure 1.3) (with the exception of two scenarios). MSEP based on error-prone covariates decreases as sensitivity and specificity increase, that is, as we introduce less error MSE improves. For both adjustment methods, MSEP increases and then decreases as sensitivity increases, for lower specificities, while for higher specificities MSE decreases as sensitivity increases.

In general, ROC-AUC was highest using the full adjustment method compared to both no adjustment and the time independent adjustment. However, ROC-AUC was slightly higher based on the error-free covariates (Figure 1.4). ROC-AUC based on the error-free covariates remained relatively constant as sensitivity and specificity varied. ROC-AUC based on error-prone covariates increases as sensitivity and specificity increase, that is, as we introduce less error ROC-AUC improves. For both adjustment methods, ROC-AUC decrease and then increase as sensitivity increases, for lower specificities, while for higher specificities ROC-AUC increases as sensitivity increases.

Overall, the full adjustment method preforms best in terms of calibration, overall model

accuracy, and discrimination.

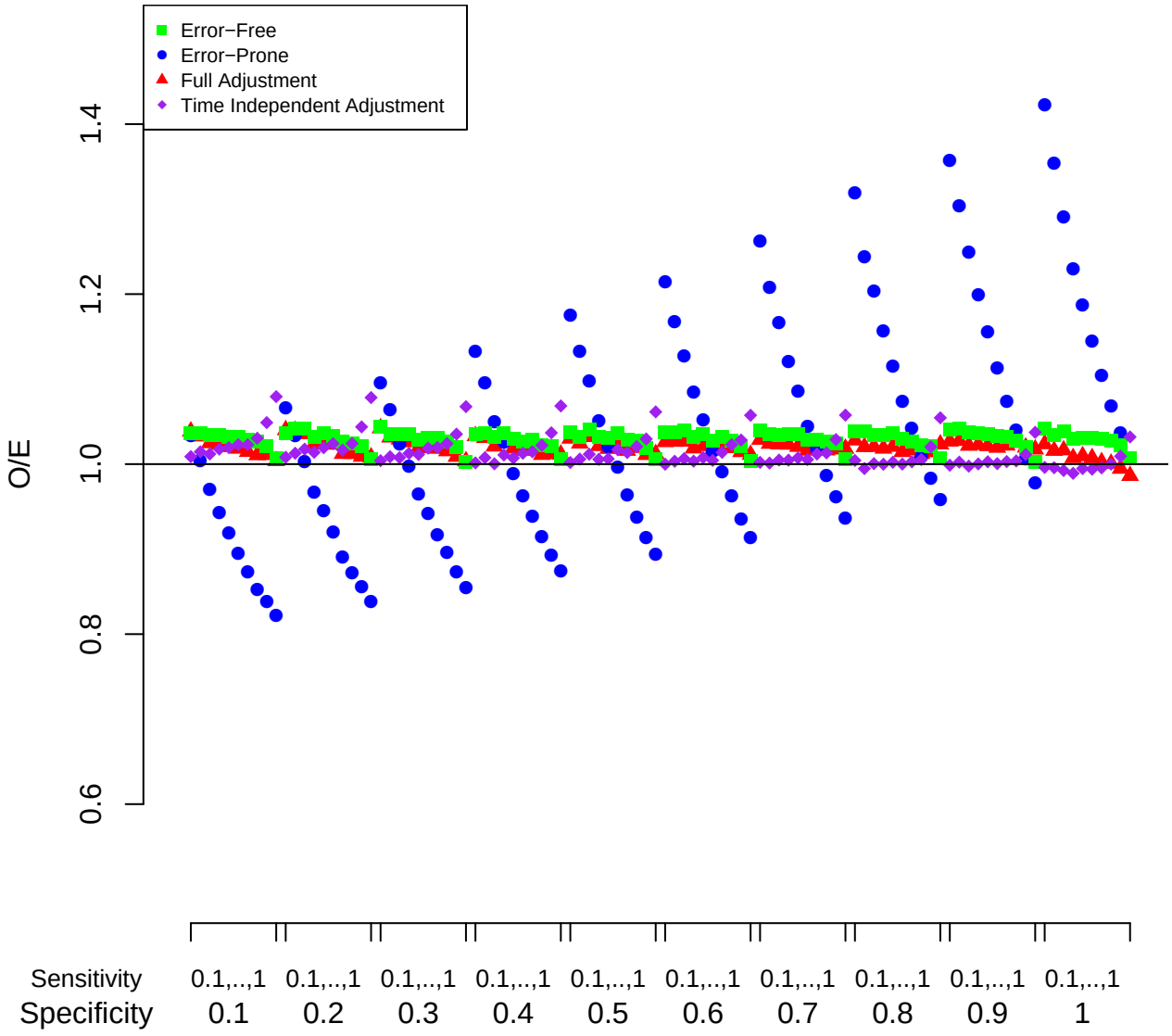


Figure 1.2: O/E ratios for PFS/OS simulations varying sensitivity and specificity.

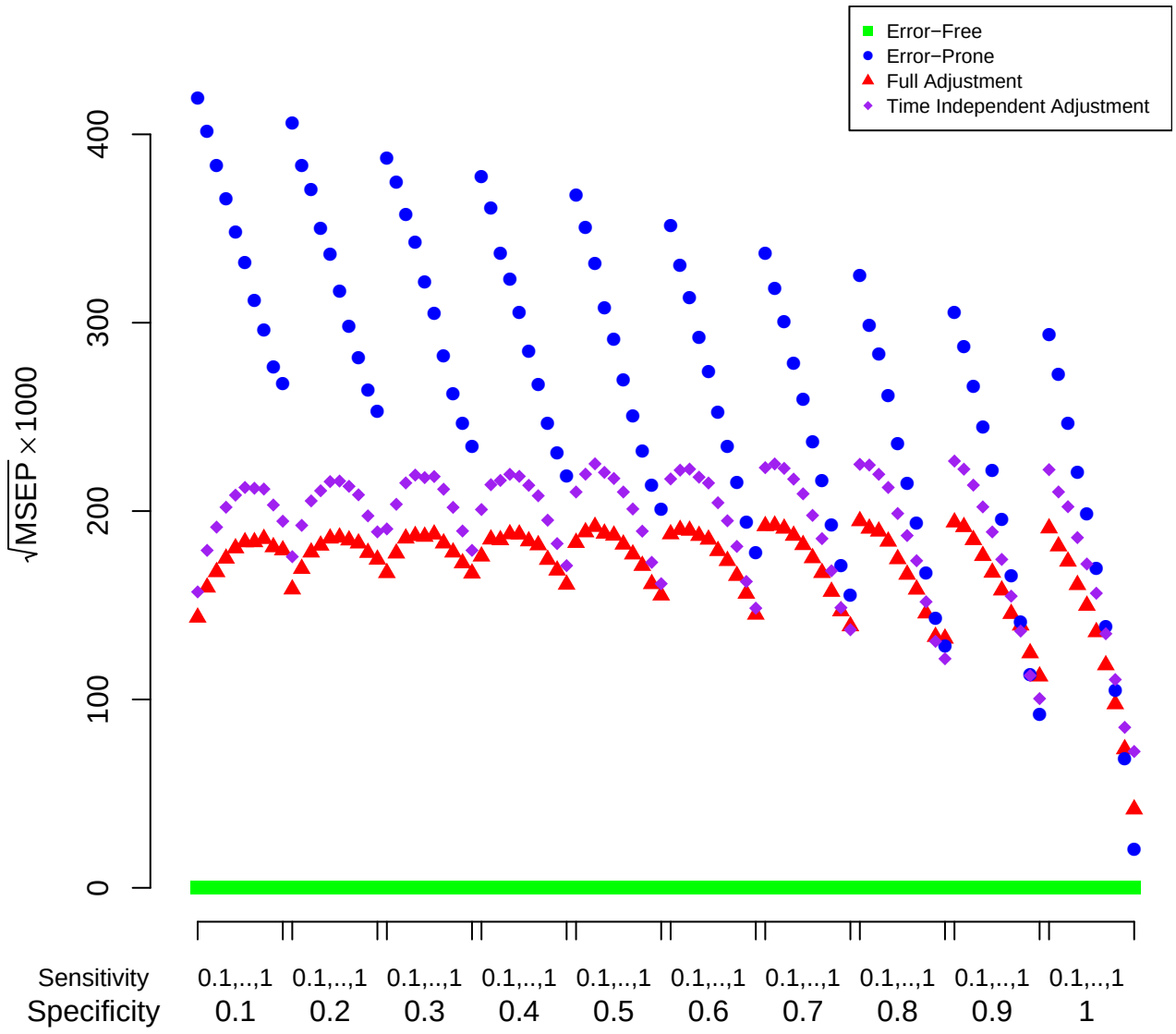


Figure 1.3: $\sqrt{MSEP} * 1000$ for PFS/OS simulations varying sensitivity and specificity.

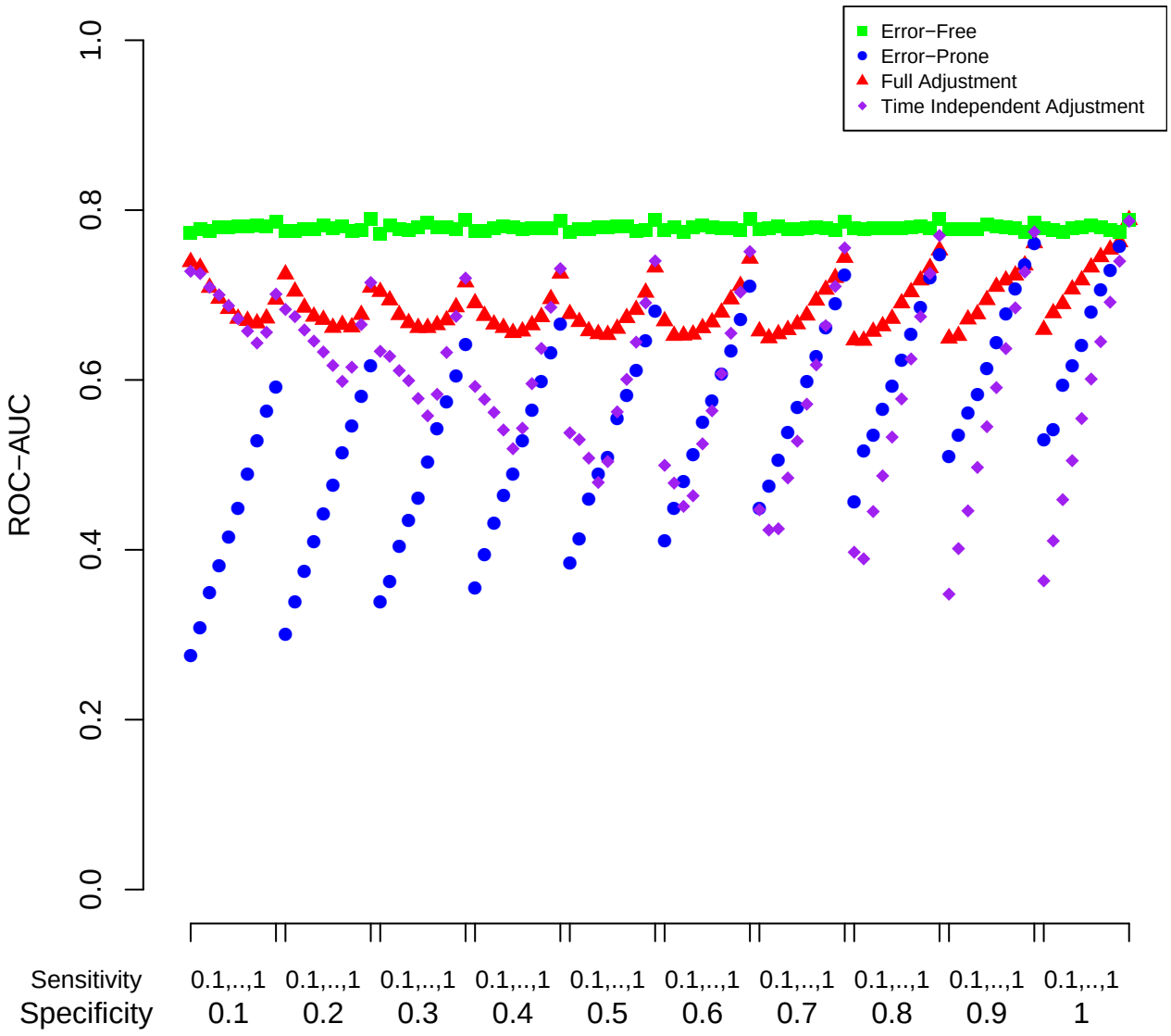


Figure 1.4: ROC-AUC for PFS/OS simulations varying sensitivity and specificity.

1.6 Data Application

We illustrate our proposed method using a data application in the context of Mendelian risk prediction models focusing only on misreporting of breast cancer. For the consultands, we use data from the Cancer Genetics Network (CGN). CGN consists of families with personal or family history of cancer. The data set includes 2,038 families with 34,310 relatives. 9.2% of relatives have breast-cancer. For this study, only error-prone, self-reported family history is available. In addition to family history, this data set also contains BRCA1/2 testing results for each consultand.

For our validation study, we use data from University of California at Irvine (UCI) (Ziogas and Anton-Culver (2003)). This study included cancer affected consultands with either breast, ovarian, or colon cancer. The data set includes 719 families with 1,521 female relatives. 19.3% of relatives have breast-cancer. Both error-free and error-prone family history are available in this data set.

We estimate the measurement error process using smoothed Kaplan-Meier estimators using the UCI validation study. We then apply our method to the CGN consultands. For each consultand, using BRCAPRO, we estimate her probability of being a carrier for a mutation given her error-prone family history, as well as given our proposed adjustment. Using the true BRCA1/2 carrier status, O/E ratios, Brier-scores, and ROC-AUC were calculated based on the error-prone family history as well as based on our proposed adjustment. The bandwidths for the smoothed Kaplan-Meir were selected so that calibration for being a BRCA, BRCA1, and BRCA2 carrier were closest to 1.

The respective O/E ratios of being a BRCA, BRCA1, and BRCA2 carrier are 1.007, 1.073, and 0.916 based on error-prone family history; and 0.976, 1.037, and 0.892 based on the adjustment. The respective Brier scores for being a BRCA, BRCA1, and BRCA2 carrier are 0.141, 0.102, 0.058 based on error-prone family history; and 0.139, 0.102, and 0.057 based on the adjustment. The respective ROC-AUC for BRCA, BRCA1, and BRCA2 carriers are 0.777, 0.791, and 0.725 based on the error-prone family history; and 0.776, 0.787, and 0.722 based on the adjustment.

Overall, we see a slight improvement in Brier score based on the adjustment. Before the

adjustment the model was well calibrated for BRCA, but not as well calibrated for BRCA1 and BRCA2 separately. The adjustment improves BRCA1 calibration, while the calibration of BRCA2 is slightly worse. ROC-AUC are slightly worse using the adjustment.

Results of O/E ratios of being a BRCA carrier stratified by risk are shown in Figure 1.5. Individuals are ordered by their probabilities of being a BRCA carrier based on the error-prone family history, and stratified into 10 strata. O/E ratios as well as 95% confidence intervals are calculated for each strata. Using error-prone family history, the model is not well calibrated in the low risk deciles. We see that the O/E ratio is greater than one in these deciles, implying that the model underestimates the risk for these individuals. Insurance companies, will often approve genetic testing only for individuals whose estimated carrier probability is above a certain cutoff, thus calibration is especially important in the low risk deciles, since clinical decisions will be made based on estimated carrier probabilities. Our proposed adjustment improves calibration in the low risk deciles by a significant amount, which will lead to better clinical decisions.

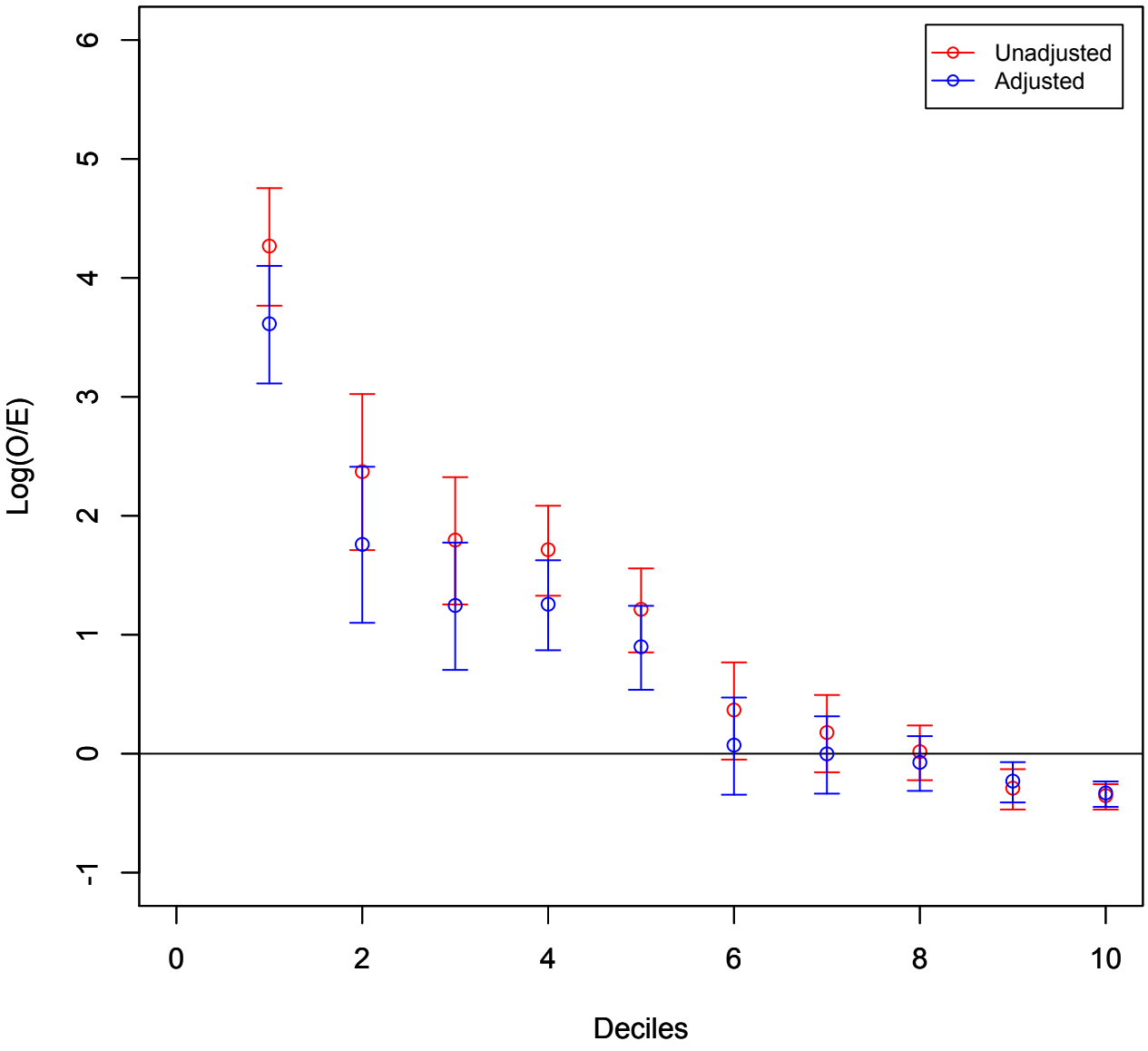


Figure 1.5: Log of observed over expected ratios and 95% confidence intervals for being a BRCA carrier for families in CGN data set stratified by risk deciles.

1.7 Discussion

In this chapter we explore a method to adjust for measurement error in time to event data. Previous literature has focused on adjusting for measurement error in survival outcomes, but not for time to event data measured with error. Our method adjusts for measurement error in time to event data used as covariates, and is applicable to both the setting of Mendelian risk prediction models and multivariate survival prediction models.

Simulations studies in both of these settings show that models based on error-prone time to event data are not well calibrated. The adjustment method improves model calibration substantially. The adjustment method also improves total accuracy by improving MSEP. ROC-AUC in Mendelian risk prediction models either remains the same or is improved using the adjustment method depending on the simulation scenario, while ROC-AUC in multivariate survival prediction models improves using the adjustment method. We also show that the adjustment method preforms better than an alternative time independent adjustment approach in the context of multivariate survival prediction models.

Our method assumes that the measurement error process is transportable from the validation data to the main study. This assumption should be given careful thought, as there may be scenarios for which this assumption will not hold for $P(H|H^*)$, but will hold for $P(H^*|H)$. In addition, the methods presented in this chapter for Mendelian risk prediction models assume only one disease, whereas Mendelian risk predication models include multiple diseases. An extension of this work to multiple diseases is presented in chapter 2.

The work presented in this chapter can be applicable to various scenarios, and was motivated by the context of Mendelian risk prediction models. Self-reported family history is often reported with error. Inaccurate reporting of family history could lead to inappropriate care Murff et al. (2004). Underreporting (false negatives) of cancer in the family, gives rise to an underestimation of cancer risk, which can result in inadequate screening and substandard treatment Murff et al. (2004). On the other hand, over-reporting (false positives) of cancer, gives rise to an overestimation of cancer risk, which can cause stress Douglas et al. (1999), unnecessary procedures and genetic testing Kerr et al. (1998); Sweet

et al. (2002); Fry et al. (1999). For these reasons, methods to adjust predictions based on self-reported family history are of clinical significance.

The method proposed in this chapter can be incorporated into the BayesMendel R package, and will be of great clinical use. Given a good validation study, Mendelian risk prediction models can automatically incorporate this adjustment, so that clinicians will be able to obtain more accurate risk predictions. In addition, we hope that this method will be incorporated by statisticians developing multivariate survival risk prediction models based on error-free time to event data, but implementing them using error-prone time to event data.

Extending Mendelian Risk Prediction Models to Handle Misreported Family History

Danielle Braun, Malka Gorfine, Hormuzd Katki, Argyrios Ziogas, Hoda Anton-Culver, , and Giovanni Parmigiani

Department of Biostatistics

Harvard School of Public Health

2.1 Introduction

Cancer is caused by genetic alterations which can either be inherited or can occur during one's lifetime. There is a lot of interest in identifying individuals who are at high risk of cancer due to inherited mutations. Many risk prediction models have been developed to address this problem. These models can be divided into three main types; empirical models, expert-based models, and Mendelian models. Empirical models estimate the probability of a proband (consultand) being a mutation carrier, by summarizing family history and using it as predictors in the models. Expert-based models use algorithms based on clinical judgment to calculate scores summarizing the risk. Mendelian models use Mendelian laws of inheritance of deleterious genetic variants to calculate the probability that the proband is a mutation carrier based on family history and known mutation prevalence and penetrance (the probability of having a disease at a certain age given the mutation status) (Katki et al., 2007; Parmigiani et al., 2007).

Mendelian risk prediction models for various cancers have previously been developed and are available as part of the BayesMendel R pacakge (Chen et al., 2004). These models include BRCAPRO for identifying individuals at high risk for breast or ovarian cancer by calculating the probabilities of germline deleterious mutations in BRCA1 and BRCA2. MMRPro for identifying individuals at high risk of Lynch Syndrome by calculating the probabilities of germline deleterious mutation of the MMR genes: MLH1, MSH2, MSH6. PancPRO for identifying individuals at high risk for pancreatic cancer by calculating the probabilities of germline mutations in CDKN2A, PRSS1, BRCA2, and STK11. The inputs required for these models are the prevalence of deleterious mutations in the general population and their penetrance, which are researched and continually updated. These models have all been validated in the literature (Berry et al., 2002; Chen et al., 2006; Wang et al., 2007).

This chapter focuses on BRCAPRO, but the methods developed could be extended to other Mendelian models. Women who carry a mutation in either BRCA1 or BRCA2 have an increased risk of developing breast and ovarian cancer. Mutations in BRCA1 and BRCA2 are rare in the general population (less than 0.2% of women have mutations

(Easton et al., 1995)), but are more common in families with high rates of breast or ovarian cancers (Ford et al., 1998; Newman et al., 1997; Weber, 1996). Women with family history of breast or ovarian cancer will seek genetic counseling in order to determine their probabilities of being a BRCA1/BRCA2 carrier (Croyle and Lerman, 1999).

Risk prediction models, and in particular Mendelian risk prediction models, are based on family history which is used to determine which patients are at high risk of cancer. Screening and even treatment strategies have been developed for these high risk individuals (Murff et al., 2004). These models rely on self-reported family history, but inaccurate reporting could lead to inappropriate care. Risk predictions based on underreported (false negatives) family history, can lead to inadequate screening and substandard treatment. On the other hand, risk predictions based on over-reported (false positives) family history, can cause stress, unnecessary procedures and genetic testing (Douglas et al., 1999; Fry et al., 1999; Kerr et al., 1998; Murff et al., 2004; Sweet et al., 2002).

Various studies have evaluated the accuracy of self-reported family history. Murff et al. (2004) provide a review of these studies. Anton-Culver et al. (1996), conducted a population-based study with 359 probands with breast cancer. The probands were interviewed over the phone, and family history was verified using a cancer registry. Kerber and Slattery (1997) conducted a case-control study of colon cancer, with 125 colon cancer cases and 206 controls. They conducted personal interviews, and family history was verified using a cancer registry. Ziogas and Anton-Culver (2003) conducted a study including 1111 families, with both population-based probands and clinic-based probands (with the clinic-based accounting for 6.3% of the families). These probands had either breast, ovarian, or colon cancer. They conducted personal phone interviews followed by self completed reports, and verified using medical records and death certificates. Verkooijen et al. (2004) conducted a population-based study with 219 probands with breast cancer, out of which 110 had at least one relative with breast cancer, 9 had at least one relative with ovarian cancer, and 100 had no relatives with breast or ovarian cancers. Family history was collected using a self-reported survey and was verified using a cancer registry. More recently, Mai et al. (2011) collected data as part of the population-based 2001 Connecticut Family Health Study. Family history for 1,019 individuals was collected on breast,

colorectal, prostate, and lung cancers through two phone interviews. The 1,019 participants reported family history for 20,578 first and second degree-relatives, of which a sample of 2,605 were validated. They were validated using state cancer registries, Medicare databases, the National Death Index, death certificates, and health-care facility records. Sensitivity rates for the accuracy of cancer status reporting of first-degree relatives vary in these studies (Table 2.1), but are higher in breast cancer compared to ovarian cancer in all studies. For the first four studies sensitivity rates are higher than 80%, and specificity rates are above 90%. Lower sensitivity rates were reported in Mai et al. (2011), probably due to the fact that this was a population based study.

Risk prediction based on reported family history that is inaccurate will also be inaccurate. Suppose you have a family illustrated in Figure 2.1. The proband in this family underreports cancer in three of the relatives (the mother, a grandmother, and cousin), and misreports the age of diagnosis for an aunt and the grandmother. Based on the reported family history (shown on the left), the probability of being a BRCA carrier is 0.0779, the probability of being a BRCA1 carrier is 0.03407, and the probability of being a BRCA2 carrier 0.04384. Whereas, based on the true family history (shown on the right), the probability of being a BRCA carrier is 0.80314, the probability of being a BRCA1 carrier is 0.59420, and the probability of being a BRCA2 carrier 0.20830. This is an extreme case of misreporting, but in general risk prediction calculations based on the reported family history may be very different than based on the true family history, and can lead to inadequate care.

Table 2.1: Studies Evaluating Sensitivity and Specificity of Reported Cancer in First-Degree Relatives

Study	Type of Cancer	Proband Affected/Healthy/General Population	Sensitivity	Specificity
Anton-Culver et al. (1996)	Breast	Affected Probands	54/59(92%)	364/370(98%)
Kerber and Slattery (1997)	Breast	Affected Probands	11/13(85%)	107/112(96%)
Kerber and Slattery (1997)	Breast	Healthy Probands	18/22(82%)	167/184(91%)
Ziogas and Anton-Culver (2003)	Breast	Affected Probands	188/197(95%)	850/873(97%)
Verkooijen et al. (2004)	Breast	Affected Probands	60/61(98%)	247/249(99%)
Mai et al. (2011)	Breast	General Population	(65%)	(99%)
Kerber and Slattery (1997)	Ovarian	Affected Probands	2/3(67%)	117/122(96%)
Kerber and Slattery (1997)	Ovarian	Healthy Probands	1/2(50%)	201/204 (99%)
Ziogas and Anton-Culver (2003)	Ovarian	Affected Probands	35/42(83%)	1017/1028(99%)
Verkooijen et al. (2004)	Ovarian	Affected Probands	4/6(67%)	168/170(99%)

Murff et al. (2004), Mai et al. (2011). Mai et al. (2011) report only sensitivity and specificity rates without providing the actual number of individuals.

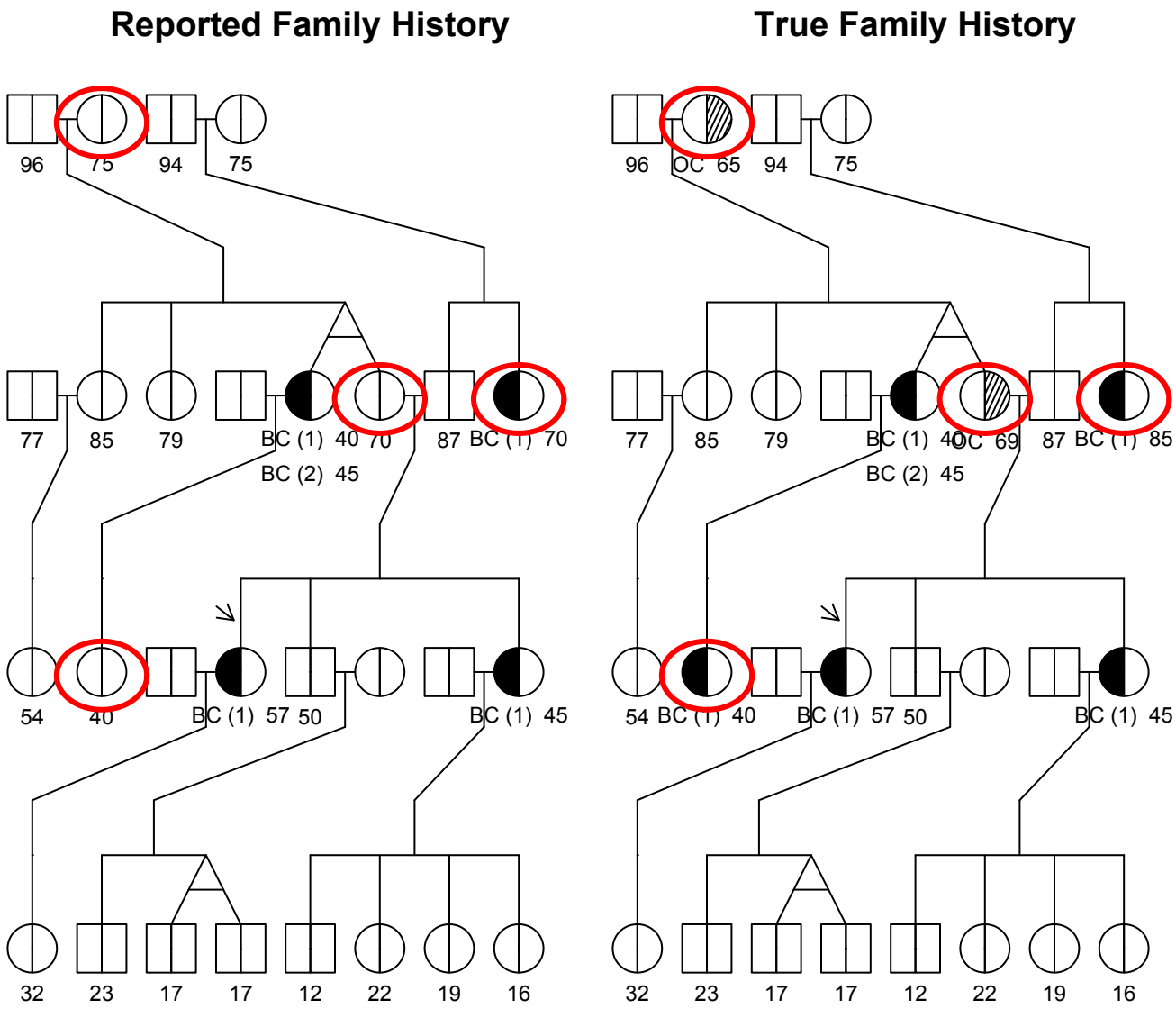


Figure 2.1: Sample Pedigree. A hypothetical example of a family with reported family history (on the left) and true family history (on the right). The proband is indicated by an arrow.

Katki (2006) studies the effect of misreported family history on Mendelian risk prediction models in more detail. Family history is composed of two components; diagnosis

(yes/no) and age-at-diagnosis (if diagnosis is yes) or current age or age of death (if diagnosis is no). Katki classifies the types of errors in the reporting of a relative's family history into three categories: (1) Diagnosis is incorrect but age-at-diagnosis is correct, (2) Diagnosis is correct but age-at-diagnosis is incorrect, (3) Both diagnosis and age-at-diagnosis are incorrect. He focuses on the first type of error, but shows the distortion caused by all three types of errors in BRCAPRO.

Katki (2006) considers underreporting of cancer disease status in three types of families: (a) proband alone (b) proband and first-degree relative (c) proband, first-degree relative, and second-degree relative. Katki derives the distortions that different scenarios of underreporting of cancer in these families would cause. Applying this to BRCAPRO, he shows that the worst error would be caused by underreporting of ovarian cancer in probands. This is seen because ovarian cancer yields higher penetrance density ratios than breast cancer. The weakest errors are underreporting of breast cancer in second-degree. In addition, Katki studies the misreporting of age of an affected relative by +/- 5 years and +/- 15 years. Considering both errors in underreporting of disease status and rounding of age, he shows that the errors in underreporting of disease status dominate.

There is extensive literature on measurement error in binary and continuous covariates (Carroll, 2006a, among others). Family history, however, is time to event data (since it includes both the disease status and age), which is used as covariates in the Mendelian risk prediction models. We are not aware of any literature directly applicable to this setting in which both the disease status and age of disease can be reported with error. A proposed method to handle this type of error is described in chapter 1. Here, we apply this adjustment method to Mendelian risk prediction models, more specifically to BRCAPRO. We begin by introducing general notation, follow by describing the adjustment method, and illustrate the proposed model extension using real data.

2.2 Methods

2.2.1 Mendelian Risk Prediction Models

Mendelian models calculate the probability of being a mutation carrier given family history. Family history has two components; disease status and age. For those who develop disease, their age will be their age of diagnosis, for those who did not develop disease their age will be their current age or age of death. Let $T_i = (T_{i1}, ..., T_{iJ})$, $T_{ij} = \min(T^o_{ij}, C_{ij})$ where T^o_{ij} is the age of disease diagnosis j for individual i, C_{ij} is the current age or age of death for individual i for disease j ,and $\delta_i = (\delta_{i1}, ..., \delta_{iJ})$ is the disease status, where $\delta_{ij} = \mathbf{1}(T^o_{ij} \leq C_{ij})$ for individual i for disease j. Let $H_i = (T_i, \delta_i)$ and $H = (H_0, H_1, ..., H_R)$ for a family of size R. Let $\gamma_i = (\gamma_{i1}, ..., \gamma_{iM})$, where $\gamma_{im} = 1$ indicates carrying the genetic variants that confer disease risk for each individual i at a gene m, and $\gamma_{im} = 0$ otherwise. For example in BRCAPRO, $M = 2$ (BRCA1 and BRCA2).

Our goal is to calculate the proband's carrier probability $P(\gamma_0|H_0, H_1, .., H_R)$. Using Bayes rule, this can be calculated as follows (Blackford and Parmigiani, 2010):

$$P(\gamma_0|H_0, H_1, .., H_R) = \frac{P(\gamma_0)P(H_0, H_1, ..., H_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0)P(H_0, H_1, ..., H_R|\gamma_0)}. \tag{2.1}$$

$P(\gamma_0)$ is the prevalence of mutation carriers in the general population. $P(H_0, H_1, ..., H_R|\gamma_0)$ is the probability of the phenotypes for the entire family given the genotype for the proband. Using Bayes rule, this probability can be rewritten as:

$P(H_0, H_1, ..., H_R|\gamma_0) = \sum_{\gamma_1,...,\gamma_R} P(H_0, ..., H_R|\gamma_0, ..., \gamma_R)P(\gamma_1, ...\gamma_R|\gamma_0)$.

$P(\gamma_1, ...\gamma_R|\gamma_0)$ are known for all genotype combinations based on Mendelian laws of inheritance. Assuming conditional independence of phenotypes given genotypes implies: $P(H_0, ..., H_R|\gamma_0, ..., \gamma_R) = \prod_{i=1}^{R} P(H_i|\gamma_i)$, where $P(H_i|\gamma_i)$, the probability of phenotype given genotype, is referred to as the penetrance. Under these assumptions we can rewrite the proband's carrier probability as:

$$P(\gamma_0|H_0, H_1, .., H_R) = \frac{P(\gamma_0)\sum_{\gamma_1,...,\gamma_R}\prod_{i=1}^{R} P(H_i|\gamma_i)P(\gamma_1, ...\gamma_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0)\sum_{\gamma_1,...,\gamma_R}\prod_{i=1}^{R} P(H_i|\gamma_i)P(\gamma_1, ...\gamma_R|\gamma_0)}. \tag{2.2}$$

2.2.2 Adjustment Method

Let H^* indicate the misreported family history; where $H_i^* = (T_i^*, \delta_i^*)$ and $H^* = (H_0, H_1^*, ..., H_R^*)$ for a family of size R. We assume the proband does not misreport his/her own history, therefore $H_0^* = H_0$. Our goal is to estimate $P(\gamma_0|H^*)$. The Mendelian risk prediction models previously developed assume true family history, since penetrance estimates for these models are based on a meta-analysis for which we expect most studies to be based on true family history. Therefore, these models calculate $P(\gamma_0|H)$. Using the total law of probability and Bayes rule, the probability of interest can be rewritten as:

$$P(\gamma_0|H^*) = \sum_H P(\gamma_0, H|H^*) = \sum_H P(\gamma_0|H^*, H)P(H|H^*) = \sum_H P(\gamma_0|H)P(H|H^*). \quad (2.3)$$

A sum is preformed in Equation (2.3) rather than an integral, since age in these models is treated as discrete with the following range; 1,...,120. The last equality in Equation (2.3) follows from the surrogacy assumption, that the carrier probability conditional on both the reported, H^*, and true, H, history is the same as the carrier probability conditional only the true history H. This assumption is plausible since this carrier probability should only be influenced by the true history. $P(\gamma_0|H)$ is then calculated using Equation (2.2), which is available as part of the BayesMendel R package. The measurement error process, $P(H|H^*)$, is estimated in a validation data set using smoothed Kaplan-Meier estimators. Since summing over all possible H is computationally intensive, we use Monte Carlo integration using $P(H|H^*)$ to select random H's.

This proposed approach assumes that the measurement error model $P(H|H^*)$ is transportable. Ideally, we would like to have a validation as similar as possible to the target population for this assumption to hold. There are scenarios for which $P(H|H^*)$ might not be transportable, but $P(H^*|H)$ is, in which case we could consider a different modeling approach, weighing the penetrance by $P(H^*|H)$. This approach is not the focus of this chapter, but is described in detail in the appendix A.

2.2.3 Measurement Error Model

We model the measurement error process $P(H|H^*)$ using validation data. Treating $H^* = (T^*, \delta^*)$ as covariates, we can model the measurement error process using a survival distribution assuming conditional independence of event and censoring times given T^*, δ^*:

$$P(T, \delta|T^*, \delta^*) = \lambda(T|\delta^*, \delta^*)^{\delta} S(T|T^*, \delta^*) h(T|T^*, \delta^*)^{1-\delta} G(T|T^*, \delta^*). \tag{2.4}$$

λ and S are the hazard and survival functions of the event time, and h and G are the hazard and survival functions of the censoring time. Depending on the data structure, one could estimate the survival distribution $P(T, \delta|T^*, \delta^*)$ parametrically (for example using a Weibull distribution or accelerated failure time model), semi-parametrically (for example using a Cox model), or non-parametrically (for example using Kaplan-Meier estimators).

We decide to use smoothed Kaplan-Meier estimators, requiring no parametric assumptions on the data. The model is stratified based on δ^*, and information is borrowed from neighborhoods based on the values of our continuous covariate T^*.

2.2.4 Measurement Error Model Involving Multiple Cancers

Family history, H, contains multiple time to event data, corresponding to the multiple diseases in the model. The measurement error process for an individual i can be written as: $P(H_i|H_i^*) = P(T_i, \delta_i|T_i^*.\delta_i^*) = P(T_{i1}, ..., T_{iJ}, \delta_{i1}, ..., \delta_{iJ}|T_{i1}^*, ..., T_{iJ}^*, \delta_{i1}^*, ..., \delta_{iJ}^*)$. We propose different approaches to model this distribution.

The first is a competing risk approach which involves taking the first event for each individual. Define the first event as $T_i^o = min(T_{i1}^o, ..., T_{iJ}^o)$, and $T_i = min(T_i^o, C_i)$, $\delta_i = \mathbf{1}(T_i^o \leq C_i)$, where $\mathbf{1}(J = j)$ indicates a failure of type j occurred. Similarly ,we can define the error-prone failure times as the first event or censoring that is reported; T_i^*, δ_i^*, and $\mathbf{1}(J^* = j^*)$ indicating a failure of type j^* was reported. We propose treating (T^*, δ^*, j^*) as covariates and modeling the measurement error process using a survival distribution

assuming conditional independence of event and censoring times given T^*, δ^*, j^*:

$$P(T_i, \delta_i | T_i^*, \delta_i^*) =$$
$$\lambda_j(T_i|T_i^*, \delta_i^*, j^*)^{\delta_i} S(T_i|T_i^*, \delta_i^*, j^*) h(T_i|T_i^*, \delta_i^*, j^*)^{1-\delta_i} G(T_i|T_i^*, \delta_i^*, j^*). \quad (2.5)$$

λ_j is the cause-specific hazard and S is the overall survival function of the event time, and *h* and *G* are the hazard and survival functions of the censoring time. We propose using smoothed Kaplan-Meier estimators. The main limitation of this proposed approach is that it only uses the first event type.

Alternatively, one could model $P(H|H^*)$, as a multivariate distribution. This could be done either by assuming independence of the measurement error across diseases:
$P(T_{i1}, ..., T_{iJ}, \delta_{i1}, ..., \delta_{iJ} | T_{i1}^*, ..., T_{iJ}^*, \delta_{i1}^*, ..., \delta_{iJ}^*) = \prod_{j=1}^{J} P(T_{ij}, \delta_{ij} | T_{ij}^*, \delta_{ij}^*)$.
Or if the independence assumption does not hold, one could model
$P(T_{i1}, ..., T_{iJ}, \delta_{i1}, ..., \delta_{iJ} | T_{i1}^*, ..., T_{iJ}^*, \delta_{i1}^*, ..., \delta_{iJ}^*)$ using a multivariate survival distribution (for example using frailty models). Our recommendation is to select an approach based on the context of the problem.

2.3 Results

2.3.1 Measurement Error Validation Data Source

We were able to obtain validation data from a family registry of probands with breast, ovarian, and colorectal cancers at the University of California at Irvine (UCI). Detailed information of the cohort characteristics can be found elsewhere (Ziogas et al., 2000). Ziogas and Anton-Culver (2003) provide details on the measurement error in this cohort. Briefly, this validation study had 1111 families which had at least one relative verified, of which 670 families had a proband affected with breast cancer, 123 families had a proband affected with ovarian cancer, and 318 families had a proband affected with colorectal cancer. Family history was collected from the proband by an initial phone interview. A verification report and pedigree was produced based on this phone interview. It was then mailed to the proband to complete items that were unknown and verify information. Family history of the relatives was verified by the following methods: (1) obtaining

pathology reports, tumor tissue samples, or clinical records (2) obtaining self-reports from the relatives themselves (3) obtaining death certificates on deceased relatives.

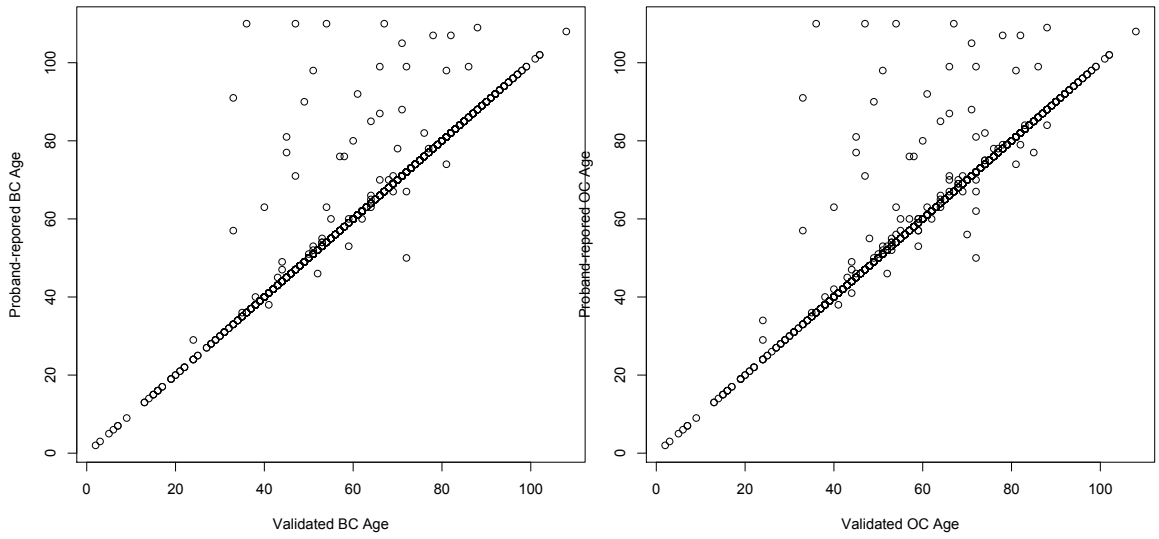
Relatives were classified into first-degree: parents, siblings, and children of the proband; second-degree: grandparents, aunts, uncles, half-siblings, nieces, and nephews of the proband; and third-degree: first cousins and grandchildren of the proband. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), for cancer disease status were calculated for various types of cancer by the degree of the relative (only rates for first and second-degree relatives were reported) (Table 2.2). In general sensitivity and PPV were lower for second-degree relatives compared to first-degree relatives, and even lower for third-degree relatives (with the exception of brain, pancreas, female breast, and leukemia). NPV and specificity were high (greater than 0.95) for all cancer types and degree of relatives. For first-degree relatives PPV and sensitivity was lowest for cancers of the female pelvic organs (pelvic and endometrium) and bladder.

Table 2.2: Misreporting of Cancer Status in UCI Study, Ziogas and Anton-Culver (2003)

Type of Cancer and Relative	Sensitivity	Specificity	PPV	NPV
Breast, 1st Degree	95.4%	97.4%	89.1%	98.9%
Breast, 2nd Degree	82.4%	97.6%	89.6%	95.8%
Ovarian, 1st Degree	83.3%	98.9%	76.1%	99.3%
Ovarian, 2nd Degree	44.1%	98.5%	62.7%	96.8%

The focus of our analysis is on breast and ovarian cancers, and the data set we obtained for this analysis included 719 families with 1,521 female relatives which were validated. The average family had 2.1 female relatives that were verified. 294 relatives (19.3%) had breast cancer, and 70 relatives (4.6%) had ovarian cancer. Ziogas and Anton-Culver (2003) focus on misclassification of disease status. However, since family history consists of age and disease status, we will consider misreporting of age as well. Figure 2.2 shows the reported age versus validated age for breast and ovarian cancer. For breast cancer we have 59 relatives (3.9%) for which the reported and validated age are not equal. For ovarian cancer we have 89 relatives (5.9%) for which the reported and validated age are not equal. Both error-free and error-prone family history are available in this data set. We fit the measurement error model on this data set separately for breast and ovarian

cancers.



a) Breast Cancer b) Ovarian Cancer

Figure 2.2: Reported versus validated ages of breast and ovarian cancers in female relatives in UCI data.

2.3.2 CGN Model Validation Study

The Cancer Genetics Network (CGN) is a national network funded by the National Cancer Institute. The data obtained for this analysis consists of families with personal or family history of cancer, and includes 2,038 families with 34,310 relatives. The average family in this data set has 16.8 relatives. 3,143 relatives (9.2%) had breast cancer, and 610 relatives (1.8%) had ovarian cancer. Only error-prone, self-reported family history is available in this data set. This data set also contains BRCA1/2 testing results for each proband, in addition to family history. We apply the measurement error adjustment to this data set.

2.3.3 Applying the Proposed Adjustment Method to the CGN Model Validation Study

We use the UCI data to fit the measurement error model using smoothed Kaplan-Meier estimators (Beran, 1981) using the prodlim R package (Gerds, 2011). We develop a mea-

surement error model for breast and ovarian cancers separately, based on Equation (2.4). For each cancer type, we stratify the model based on δ^*, and estimate survival and hazards for the failure time as well as censoring time distribution. Thus, four Kaplan-Meier estimators were obtained for each cancer type. For each of these estimators, a bandwidth is required. We select the optimal bandwidth by examining a grid of bandwidths. We examined a four-dimensional grid varying each bandwidth from 0.1 to 0.9 by increments of 0.2, and estimating the measurement error distribution given each quadruplets of bandwidths.

In order to select the optimal bandwidths, 50% of the CGN data was sampled and used to select the optimal bandwidths in terms of overall calibration of being a BRCA, BRCA1, and BRCA2 carrier. This sampling process was repeated ten times. For this data application, three different sets of four-dimensional bandwidths were selected in these ten iterations. In five out of the ten iterations the same four-dimensional set of bandwidths were selected. In the remaining five iterations, one set of four-dimensional bandwidths was selected three times, and one set of four-dimensional bandwidths was selected twice. The final set of four-dimensional bandwidths was the most frequently selected set, which was selected in half of these iterations.

BRCA testing results were available for all probands in this data set. The performance of the measurement error adjustment was evaluated using three different measures. The first, evaluating model calibration by looking at the observed over expected ratios (O/E). The second, evaluating the overall fit by looking at Brier scores. The third, evaluating model discrimination by looking at the area under the receiver operating characteristic curve (ROC-AUC). We also compare our proposed approach to an alternative approach correcting for measurement error assuming measurement error is independent of age. In other words, $P(T, \delta|T^*, \delta^*) = P(\delta|\delta^*)$. We estimate $P(\delta|\delta^*)$ (NPV and PPV) in the UCI validation study, and use it to adjust for measurement error as follows;

$$P(\gamma_0|T^*, \delta^*) = \sum_T \sum_\delta P(\gamma_0|T, \delta) P(\delta|\delta^*). \tag{2.6}$$

We will refer to this as the age independent approach.

Results of these measures based on error-prone family history compared to the full adjustment method and age independent adjustment method are shown in Table 2.3. The full adjustment method improves the O/E ratio for BRCA1 (1.0113 vs. 1.0734) and BRCA2 (0.9318 vs. 0.9157) although the O/E ratio for BRCA is worse (0.9728 vs. 1.0070). Brier scores are improved for BRCA (0.1393 vs. 0.1409) and BRCA2 (0.0570 vs. 0.0583), but not for BRCA1 (0.1021 vs. 0.1019). ROC-AUC is lower for BRCA, BRCA1, and BRCA2, using the full adjustment compared to based on error-prone family history. For the age independent approach, the O/E ratio is worse for BRCA (1.042 vs. 1.0070), BRCA1 (1.2857 vs. 1.0734), and BRCA2 (0.9136 vs. 0.9157) (Table 2.3). Brier scores are slightly higher using this approach for BRCA, BRCA1, and BRCA2, and ROC-AUC using this approach are lower for BRCA, BRCA1, and BRCA2.

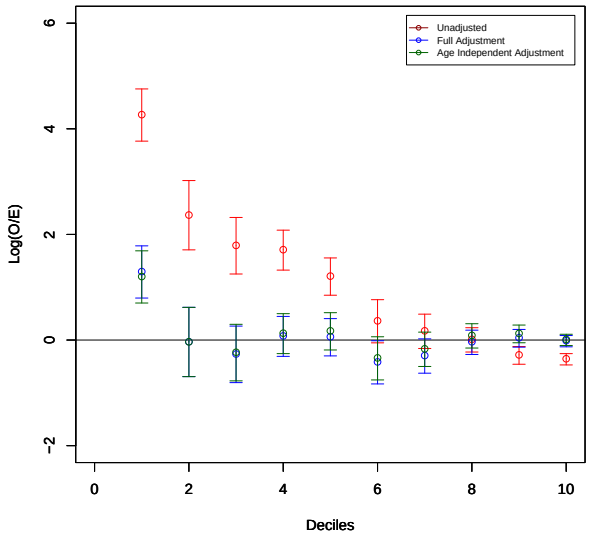
Table 2.3: Summary of results for CGN families applying proposed methods to adjust for measurement error

		O/E	ROC-AUC	Brier Score
BRCA	Error-Prone	1.0070	0.7769	0.1409
	Proposed Adjustment	0.9728	0.7424	0.1393
	Proposed Age Independent Adjustment	1.0421	0.7377	0.1401
BRCA1	Error-Prone	1.0734	0.7906	0.1019
	Proposed Adjustment	1.0113	0.7544	0.1021
	Proposed Age Independent Adjustment	1.2857	0.7544	0.1026
BRCA2	Error-Prone	0.9157	0.7244	0.0583
	Proposed Adjustment	0.9318	0.6907	0.0570
	Proposed Age Independent Adjustment	0.9136	0.6817	0.0574

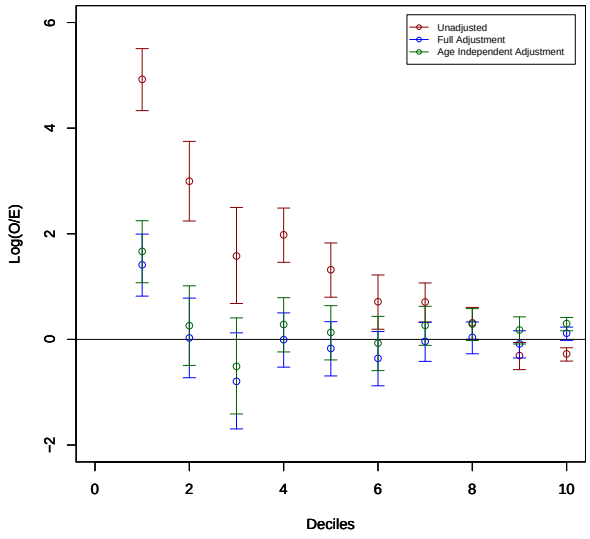
Overall, the full adjustment preforms better than the age independent adjustment in terms of calibration and Brier score. It is important to note that model calibration for BRCA based on error-prone family history was 1.0070 to begin with. This implies that this data might not have strong rates of misreporting.

Figure 2.3 shows the log(O/E) stratified by risk deciles for BRCA, BRCA1, and BRCA2. Individuals were stratified into ten strata based on their probabilities of being a carrier given their error-prone family histories. The $\log$ of the O/E ratio was calculated within each strata along with 95% confidence intervals. BRCAPRO based on error-prone family history preforms poorly in the low deciles (corresponding to large O/E ratios). Both the

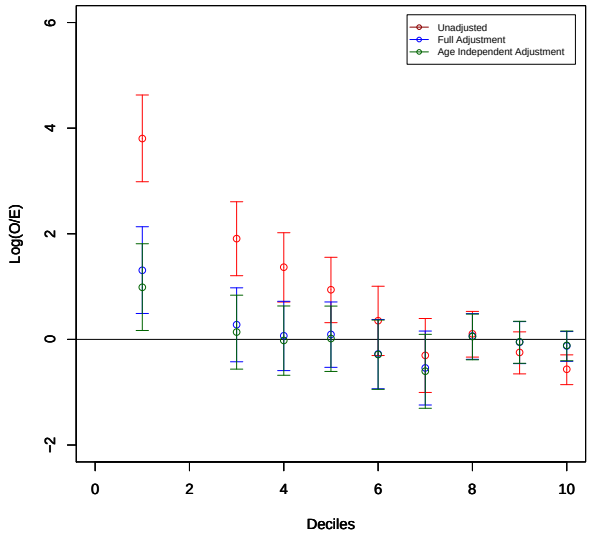
full adjustment approach and age independent approach improve calibration especially in these low decile groups, by lowering the O/E ratios. This is of great clinical significance, since individuals in these deciles often face the clinical decision of whether or not to get tested for genetic mutations. Insurance companies often use cutoffs based on risk calculations to determine which individuals would be covered for genetic testing. Individuals in high risk deciles would qualify for testing, however individuals in low risk deciles might not qualify depending on their risk. Thus, it is especially important to have a well calibrated model for these individuals. The fact that O/E ratios in the low risk deciles are greater than one implies that BRCAPRO underestimates the risk for these individuals. This is of great clinical concern, as some individuals who might be carriers are not tested due to this underestimation. Both adjustments improve model calibration, but the full adjustment preforms better than the age independent adjustment for BRCA1 in the high risk quantiles.



a) BRCA



b) BRCA1



c) BRCA2

Figure 2.3: Log(O/E) and 95% confidence intervals for CGN families stratified by risk.

2.3.4 Simulated Families

Extensive simulations have been conducted in chapter 1. Since data sources containing both error-free and error-prone family histories are limited, we decide to illustrate our

method on simulated families. These families were generated to have similar characteristics to the families in the CGN data set. We introduce error in disease status using sensitivity=0.649 and specificity=0.990 for breast cancer taken from Mai et al. (2011), and sensitivity=0.833 and specificity=0.989 for ovarian cancer taken from Ziogas and Anton-Culver (2003). We introduce error in age assuming an additive classical model; $T^* = T + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, and $\sigma = 3$. We use the UCI data to fit the measurement error model for these simulations. The bandwidth was selected to minimize mean squared error in predictions.

Figure 2.4 compares predictions based on error-free, error-prone, and the full adjustment method in this simulation setting. Mean squared error in prediction was 0.02726 based on error-prone family history and 0.02620 based on the full adjustment. Results for being a BRCA carrier are shown in the top panel in red, BRCA1 carrier in the middle panel in purple, and BRCA2 carrier in the bottom panel in green. The left column shows predictions based on error-free family history compared to error-prone family history. In general there is more underreporting of cancer (individuals whose carrier probabilities are below the 45^o line) due to the low sensitivity in this data. The middle column shows predictions based on error-free family history compared to the full adjustment method. For BRCA1 we have more individuals whose predictions based on the adjustment method are higher than based on the error-free compared to BRCA2. The right column shows predictions based on error-prone family history compared to the full adjustment method. For BRCA and BRCA1 we see that the full adjustment increases the carrier probabilities for low risk individuals and decreases carrier probabilities for high risk individuals, whereas for BRCA2 the adjustment is not as strong. This is likely due to the fact that ovarian cancer, which has high rates of misreporting in the UCI data, affects the BRCA1 carrier status more than the BRCA2 carrier status.

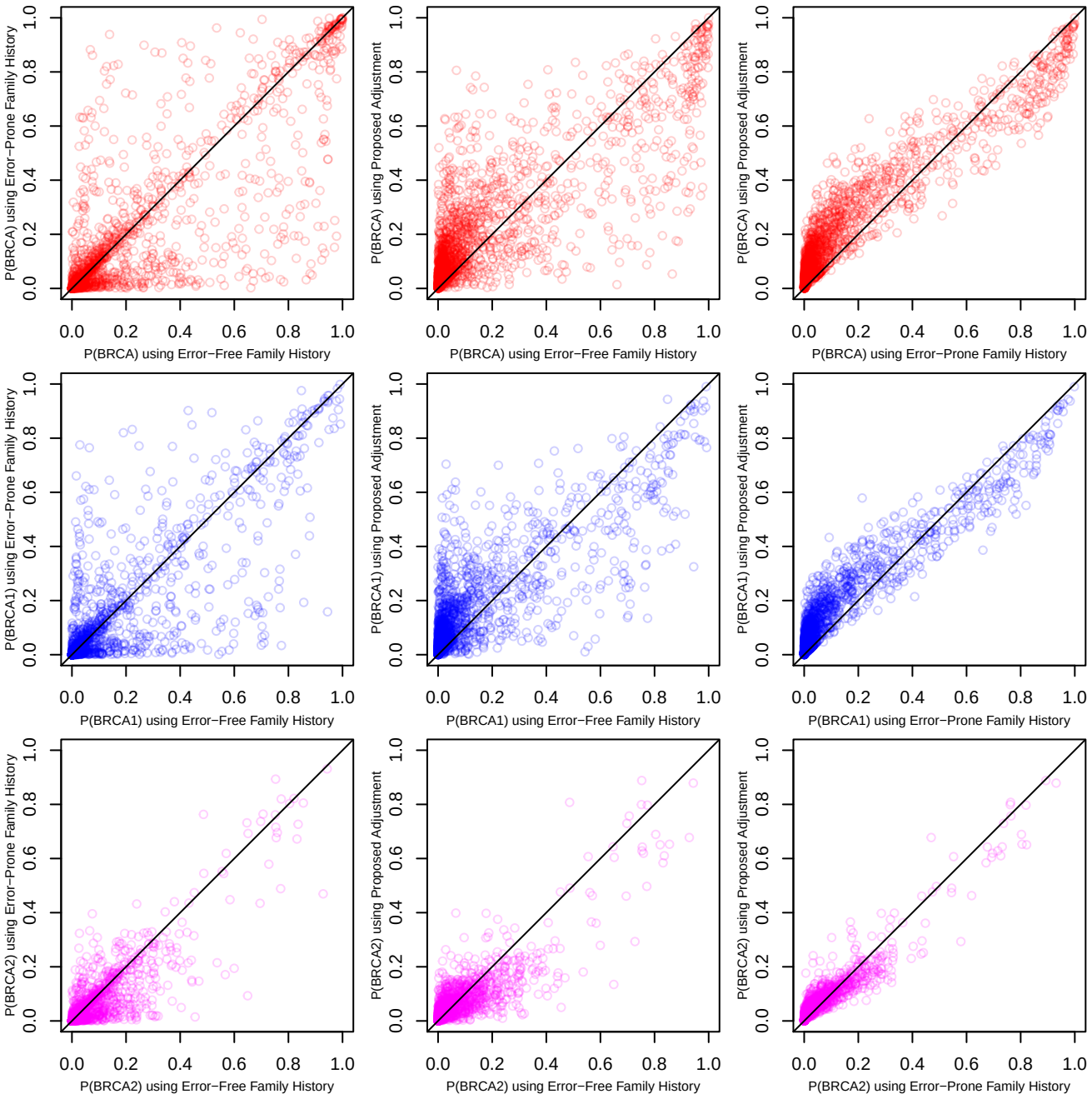


Figure 2.4: Carrier probabilities based on simulations comparing probabilities based on error-free, error-prone, and the proposed adjustment method. BRCA carrier probabilities (in red), BRCA1 carrier probabilities (in purple), and BRCA2 carrier probabilities (in green).

In addition, we compare the full adjustment to an age independent adjustment using PPV and NPV estimates based on the UCI data. Mean squared error in prediction was 0.02726

based on error-prone family history and 0.04803 based on the age independent adjustment. The age independent method does not preform well in this simulation setting.
Overall, the full adjustment method improves predictions by increasing predictions for low risk individuals and decreasing predictions for high risk individuals.

2.4 Discussion

In this chapter we propose a method to adjust for misreporting of family history in Mendelian risk prediction models. Previous literature has focused on evaluating misclassification of disease status for various cancers by comparing proband reported family history to various gold standards. Katki (2006) studies the effects of misreporting of family history on Mendelian risk prediction models. In this chapter, for the first time, we implement a measurement error adjustment in Mendelian risk prediction models.
We compare two methods to adjust for measurement error in the reporting of two cancers (breast and ovarian). The first using non-parametric smoothed Kaplan-Meier estimators, and the second using an age independent approach (using PPV and NPV estimates). Both in the context of the data application and in simulations, the full adjustment outpreforms the age independent adjustment. Using the full adjustment, we see improved calibration in BRCA1 and BRCA2, especially in low risk individuals.
It should be noted that one limitation of the UCI data is that it included only affected probands, meaning that the rates of misreporting might not be generalizable. Also, only a subset of the relatives for a given family were verified. For example, the average number of first-degree relatives for a proband with breast cancer was 6.8 out of which only 1.8 were verified, similarly the average number of second-degree relatives was 18.9 out of which only 2.3 were verified (these numbers are similar for families with probands having ovarian or colorectal cancer). Another limitation is that the data obtained represents only a subset of the families. More specifically these are families for which verification was done before July 2000. This might be a limitation since families who participated in the study earlier might be more likely to have a family history of cancer (Ziogas and Anton-Culver, 2003).

The CGN data on which we illustrate our method also has some limitations. The risk prediction model is well calibrated for being a BRCA carrier based on error-prone family history, implying that there might not be a lot of error in this data set. This data set consists of families with personal or family history of cancer. These families might be more aware of their family history, and therefore have lower rates of misreporting.

Self-reported family history is often reported with error. Inaccurate reporting of family history can lead to inappropriate care. For these reasons, methods to adjust predictions based on self-reported family history are of clinical significance. Insurance companies use fixed cutoffs to determine which patients can receive genetic testing. Without the measurement error adjustment, BRCAPRO is not well calibrated for low risk individuals. Our proposed adjustment improves calibration in this subpopulation. This will lead to better, more accurate clinical decisions for these individuals.

Given a good validation study for measurement error, Mendelian risk prediction models, such as BRCAPRO, can be extended to incorporate the proposed measurement error adjustment. This will allow clinicians and patients to obtain more accurate risk prediction estimates.

Adjustment for Mismeasured Exposure using Validation Data and Propensity Scores

Danielle Braun, Malka Gorfine, Corwin Zigler, Francesca Dominici, and Giovanni Parmigiani

Department of Biostatistics

Harvard School of Public Health

3.1 Introduction

There is a lot of interest in estimating causal treatment effects. Ideally, randomized control studies would be used to study treatment effects, but these are not always available due to ethical reasons, feasibility reasons such as cost, time constraints, and compliance (among other reasons). Observational studies are often more widely available, however, involve some limitations. Since subjects are not randomized by treatment, their characteristics might not be balanced by treatment group. In order to overcome this limitation, propensity score based methods have been proposed (Rosenbaum and Rubin, 1983).

The propensity score is defined as the probability that an individual has been assigned to treatment given their covariates. Various propensity score methods have been introduced. Rosenbaum and Rubin (1984) introduce a method that stratifies individuals based on their propensity score, and takes the average treatment effects across strata. Propensity scores can also be used to weigh individual observations (Rosenbaum, 1987). Matching individuals by their propensity scores attempts to create treated and control groups that have similar covariate values. Propensity scores have also been used as covariates in the outcome model.

We focus on the setting in which covariates are not balanced by treatment assignment, and propensity score methods are used to overcome this limitation. Although one could run a regression analysis, including all confounders in the outcome model to overcome this limitation, propensity scores are advantageous in two main settings. The first, if there are many confounders, and fewer than eight events are observed per confounder (Cepeda et al., 2003). Propensity scores allow for a way to reduce the dimensionality, and preform better than standard regression. The second, in the case of model misspecification, standard regression will lead to bias. It has been shown that treatment effect estimates are more sensitive to outcome model misspecification compared to an incorrect propensity score model (Drake, 1993).

Propensity score methods assume that treatment assignment is measured without error, but in reality treatment assignment in observational studies could be measured with error. We came across this in comparative effectiveness research, where accurate procedural

codes are not always available. More specifically, our work is motivated by a comparative effectiveness research study assessing the use of surgery to remove a brain tumor in an elderly population diagnosed with glioblastoma. We are able to obtain Medicare Part A data, which is a large data set (41,971 individuals) containing information on our outcome of interest (mortality), the treatment assignment (surgery), as well as many confounders. However, treatment assignment in this data set, is based on ICD9 billing codes which are not always an accurate measure of the treatment. For a subset of individuals (5,463 individuals), we are able to obtain data from SEER-Medicare. For these individuals, we have more accurate treatment assignment information. The SEER-Medicare is our internal validation study.

Treatment assignment in this context, can be thought of as the exposure variable. Measurement error in exposures has been extensively studied in the measurement error literature. Various techniques to adjust for measurement error in this setting have been developed including likelihood based approaches, regression calibration, Bayesian approaches, among others (Carroll, 2006b). A literature review conducted by Jurek et al. (2006), shows that measurement error in exposures is often ignored in studies. Causal inferences about the effect of an exposure may be biased by errors in the exposure (Hernán and Cole, 2009). Thus, there is a clear need to adjust for measurement error in this setting. Standard techniques adjusting for measurement error in exposures cannot be applied directly to this setting. Misclassification of treatment assignment will lead to both error in the exposure variable directly, as well as error in the propensity score estimates. Previous literature using propensity score methods has focused on measurement error in confounders and missing confounders but not on measurement error in exposures. Various approaches have been proposed in this setting. Stürmer et al. (2005), propose a regression calibration approach using validation data to estimate an adjusted propensity score in the case of unmeasured confounders. McCandless et al. (2012) propose a flexible Bayesian procedure to adjust for missing confounders using external validation data. McCaffrey et al. (2013) provide a consistent inverse probability weighting estimator in the case where confounders are measured with error. We are not aware of any literature addressing measurement error in treatment assignment in this context.

We propose a two step likelihood approach. First adjusting for measurement error in the propensity score using validation data. Next, we use the adjusted propensity score and adjust for the measurement error in the treatment effect using external validation data. We compare our proposed approach using four different propensity score based methods; stratification, inverse probability weighing of the likelihood, matching, and covariate adjustment. In addition, we propose a way to eliminate bias caused by the misclassification of treatment to the inverse probability weighting (IPW) estimator directly.

In Section 3.2, we introduce general notation and formulate our model. Afterwards, in Section 3.3 we describe our proposed likelihood adjustment using various propensity score techniques. In section 3.4, we propose a way to eliminate the bias in the IPW estimator. We perform simulations in Section 3.5, and summarize the main results in Section 3.6.

3.2 General Notation and Model Formulation

3.2.1 Notations

Let Y denote the true outcome (ex: binary disease status, or a continuous measure, etc), let X denote a true binary exposure (ex: claims data medication use), let X^* denote the error-prone exposure (ex: self-reported medication use), let $\mathbf{C}$ denote a vector of confounders measured without error (ex: age, etc).

Let $i = 1, ..., N_m$ index individuals in the main study. For these individuals $Y, X^*, \mathbf{C}$ is available. In addition, suppose we have a validation study with $j = 1, ..., N_v$ individuals. For these individuals $X, X^*, \mathbf{C}$ is available. The validation study could be either an external validation study, for which Y might or might not be known, or an internal validation study, in which case Y would be known.

3.2.2 Model Formulation

Outcome Model

In general, we assume the outcome model is a generalized linear model. The true, error-free, outcome model can be written as $E(Y|X, C) = g^{-1}(\beta_{0x} + \beta_{1x}X + \beta_{2x}\mathbf{C})$, where g is known. The error-prone outcome model can be written as $E(Y|X^*, C) =$

$g^{-1}(\beta_{0x^*} + \beta_{1x^*}X^* + \beta_{2x^*}\mathbf{C})$. In the $logit$ case, which is the outcome model chosen for simulations in this paper, the true outcome model can be written as: $P(Y = 1|X, \mathbf{C}) = \frac{1}{1+e^{-(\beta_{0x}+\beta_{1x}X+\beta_{2x}\mathbf{C})}}$ and the error-prone outcome model can be written as: $P(Y = 1|X^*, \mathbf{C}) = \frac{1}{1+e^{-(\beta_{0x^*}+\beta_{1x^*}X^*+\beta_{2x^*}\mathbf{C})}}$.

Propensity Score Model

The true propensity score model can be written as: $PS_{true} = logit(P(X = 1|\tilde{\mathbf{C}})) = \gamma_{\mathbf{x}}\tilde{\mathbf{C}}$, , where $\tilde{\mathbf{C}} = (1, \mathbf{C})$. The error-prone propensity score model can be written as: $PS_{ep} = logit(P(X^* = 1|\tilde{\mathbf{C}})) = \gamma_{\mathbf{x}^*}\tilde{\mathbf{C}}$.

Measurement Error Model

The measurement error model can be written as $E(X^*|X, C) = h^{-1}(\eta_0+\eta_1X+\eta_2\mathbf{C})$, where h is known.

3.3 Likelihood Adjustment Approach

Our interest is in estimating the true treatment effect. Confounders in the data might not be balanced by treatment group. We focus on scenarios for which propensity score methods outpreform standard regression models. These scenarios include having many confounders in the model (fewer than eight events per confounder), and model misspecification of the outcome model.

We consider using different propensity score methods; stratification, weighted likelihood, matching and covariate adjustment. For each method, we propose a two step likelihood correction approach to correct for measurement error. First, use a likelihood approach to correct for measurement error in the propensity score model and estimate an adjusted propensity score. Next, use a likelihood approach on the outcome model to adjust for measurement error in the misclassified exposure variable directly using the adjusted propensity score.

3.3.1 Correcting for Measurement Error in Propensity Score

The likelihood of the propensity score model can be written as the product of the likelihood in the main study and in the validation study. In the main study, only X^* is observed. The likelihood in the main study can be rewritten using the law of total probability by summing over all possible values of the true X and multiplying by the measurement error $P(X^*|X, \mathbf{C})$. The overall likelihood (Equation (3.1)) is then maximized to obtain maximum likelihood estimates for γ.

$$\begin{aligned} L(\gamma) &= \prod_i^{N_m} P(x_i^*|\mathbf{c}_i, \gamma) \prod_j^{N_v} P(x_j|\mathbf{c}_j, \gamma) = \prod_i^{N_m} \sum_x P(x_i^*|x, \mathbf{c}_i, \gamma) P(x|\mathbf{c}_i, \gamma) \prod_j^{N_v} P(x_j|\mathbf{c}_j, \gamma) \\ &= \prod_i^{N_m} \sum_x P(x_i^*|x, \mathbf{c}_i) P(x|\mathbf{c}_i, \gamma) \prod_j^{N_v} P(x_j|\mathbf{c}_j, \gamma) \end{aligned} \tag{3.1}$$

The second equality follows from the assumption that $P(x_i^*|x, \mathbf{c}_i, \gamma) = P(x_i^*, \mathbf{c}_i|x)$, this is reasonable since we condition on both x and c. $P(X^*|X, \mathbf{C})$ is estimated using the validation data with N_v individuals. Equation (3.1) is maximized to obtain maximum likelihood estimates $\hat{\gamma}$. Using $\hat{\gamma}$ we can obtain an adjusted propensity score, $\widehat{PS_{adj}} = \hat{\gamma}\mathbf{C}$. Assuming the propensity score follows a $logit$ model we get:

$$\begin{aligned} L(\gamma) &= \prod_i^{N_m} [P(x_i^* = 1|x_i = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\gamma \mathbf{c}_i))} \\ &+ P(x_i^* = 1|x_i = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\gamma \mathbf{c}_i))})]^{x_i^*} \\ &* [P(x_i^* = 0|x_i = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\gamma \mathbf{c}_i))} \\ &+ P(x_i^* = 0|x_i = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\gamma \mathbf{c}_i))})]^{(1-x_i^*)} \\ &* \prod_j^{N_v} [\frac{1}{1 + \exp(-(\gamma \mathbf{c}_j))}]^{x_j} [1 - \frac{1}{1 + \exp(-(\gamma \mathbf{c}_j))}]^{(1-x_j)} \end{aligned} \tag{3.2}$$

3.3.2 Correcting for Measurement Error in Outcome Model

The propensity score model has now been corrected, but the outcome model still contains the misclassified treatment assignment. We use a likelihood approach on the outcome model to correct for this misclassification directly. The likelihood is different for each of the propensity score methods, therefore we illustrate this for each method individually. For each method, we illustrate the approach for both the case of an internal validation data set (as in our motivating data application), and in the case of an external validation data set (for which Y may or may not be known).

Stratification

In this propensity score method, individuals are stratified by their propensity scores into K groups with N_K individuals in each group. Treatment effect estimates are estimated for each of the K groups and averaged to obtain the overall treatment effect estimate. Ignoring the measurement error, we would stratify individuals by PS_{ep} and estimate treatment effects in each of the K strata using X^* in the outcome model. The proposed adjustment includes two steps; first stratifying individuals into K groups by PS_{adj} instead of PS_{ep}. Next, we rewrite the likelihood for the outcome model in each of the K strata using the law of total probability by summing over all possible values of the true X and weighing by the measurement error $P(X|X^*, \mathbf{C})$ which is estimated in the validation study (Equation (3.3) in the case of a validation study with known Y, and Equation (3.4) in the case of a validation study with unknown Y). We maximize $\beta_{\mathbf{k}}$ in each strata, to obtain treatment effect estimates for each strata. The overall treatment effect is the average of treatment effects across the strata.

In the case of an internal validation study, or an external validation study for which Y is

known, the likelihood can be written as follows.

$$\begin{aligned} L^*(\beta_k) &= \prod_{i \in N_k} P(y_i|x_i^*, \mathbf{c}_i, \beta_k) \prod_{j \in N_{vk}} P(y_j|x_j, \mathbf{c}_j, \beta_k) \\ L(\beta_k) &= \prod_{i \in N_k} \sum_x P(y_i|x, x_i^*, \mathbf{c}_i, \beta_k) P(x|x_i^*, \mathbf{c}_i, \beta_k) \prod_{j \in N_{vk}} P(y_j|x_j, \mathbf{c}_j, \beta_k) \\ &= \prod_{i \in N_k} \sum_x P(y_i|x, \mathbf{c}_i, \beta_k) P(x|x_i^*, \mathbf{c}_i) \prod_{j \in N_{vk}} P(y_j|x_j, \mathbf{c}_j, \beta_k) \end{aligned} \tag{3.3}$$

In the case of an external validation study, for which Y is unknown, the likelihood can be written as follows.

$$\begin{aligned} L^*(\beta_k) &= \prod_{i \in N_k} P(y_i|x_i^*, \mathbf{c}_i, \beta_k) \\ L(\beta_k) &= \prod_{i \in N_k} \sum_x P(y_i|x, x_i^*, \mathbf{c}_i, \beta_k) P(x|x_i^*, \mathbf{c}_i, \beta_k) \\ &= \prod_{i \in N_k} \sum_x P(y_i|x, \mathbf{c}_i, \beta_k) P(x|x_i^*, \mathbf{c}_i) \end{aligned} \tag{3.4}$$

The last equality follows from two assumptions, the first that $P(x_i|x_i^*, \mathbf{c}_i, \beta_k) = P(x_i|x_i^*, \mathbf{c}_i)$, which is reasonable since we condition on both X and $\mathbf{C}$. The second, that X is a surrogate for X^*, in other words that $P(y_i|x, x_i^*, \mathbf{c}_i, \beta_k) = P(y_i|x, \mathbf{c}_i, \beta_k)$. This is a very common assumption in the measurement error field.

$L(\beta_k)$ is maximized to obtain maximum likelihood estimates $\hat{\beta}_k$. We estimate the treatment effects within each strata, and report an average treatment effect estimate.

The likelihood in the case of a $logit$ outcome model is shown in the Appendix, Equation (B.1).

Weighted Likelihood

A weighted likelihood approach involves weighing each treated individual by the inverse of their propensity score and each untreated individual by the inverse of one minus their propensity score. β estimates are then obtained by maximizing the weighted likelihood. Ignoring the measurement error, each individual would be weighted according to their error-prone treatment X^* using PS_{ep}. In other words, the decision of whether to use inverse PS_{ep} or inverse $1-PS_{ep}$ as weights would be based on the error-prone treatment X^*.

The proposed adjustment in this setting includes two steps; first using PS_{adj} as weights in the likelihood. Next, the likelihood is rewritten with the adjusted weights using the law of total probability by summing over all possible values of the true X and weighing by the measurement error $P(X|X^*, \mathbf{C})$ (Equation (3.5) in the case of a validation study with known Y, and Equation (3.6) in the case of a validation study with unknown Y). We maximize $L(\beta)$ to obtain maximum likelihood estimates $\hat{\beta}$ and estimate the treatment effects. We note as a possible limitation, that the proposed adjustment does not classify the use of inverse PS_{adj} or inverse $1 - PS_{adj}$ based on the true value X, but rather based on the error-prone treatment X^*.

In the case of an internal validation study, or an external validation study for which Y is known, the likelihood can be written as follows.

$$\begin{aligned}
L^*(\beta) &= \prod_i^{N_m} [P(y_i|x_i^*, \mathbf{c}_i, \beta)]^{\frac{1}{P(x_i^*|\mathbf{c}_i, \gamma_{x^*})}} \prod_j^{N_v} [P(y_j|x_j, \mathbf{c}_j, \beta)]^{\frac{1}{P(x_j|\mathbf{c}_j, \gamma_x)}} \\
L(\beta) &= \prod_i^{N_m} [\sum_x P(y_i|x, x_i^*, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i, \beta)]^{\frac{1}{P(x_i^*|\mathbf{c}_i, \gamma_{adj})}} \prod_j^{N_v} [P(y_j|x_j, \mathbf{c}_j, \beta)]^{\frac{1}{P(x_j|\mathbf{c}_j, \gamma_x)}} \\
&= \prod_i^{N_m} [\sum_x P(y_i|x, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i)]^{\frac{1}{P(x_i^*|\mathbf{c}_i, \gamma_{adj})}} \prod_j^{N_v} [P(y_j|x_j, \mathbf{c}_j, \beta)]^{\frac{1}{P(x_j|\mathbf{c}_j, \gamma_x)}} \qquad (3.5)
\end{aligned}$$

In the case of an external validation study, for which Y is unknown, the likelihood can be written as follows.

$$\begin{aligned}
L^*(\beta) &= \prod_i^{N_m} [P(y_i|x_i^*, \mathbf{c}_i, \beta)]^{\frac{1}{P(x_i^*|\mathbf{c}_i, \gamma_{x^*})}} \\
L(\beta) &= \prod_i^{N_m} [\sum_x P(y_i|x, x_i^*, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i, \beta)]^{\frac{1}{P(x_i^*|\mathbf{c}_i, \gamma_{adj})}} \\
&= \prod_i^{N_m} [\sum_x P(y_i|x, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i)]^{\frac{1}{P(x_i^*|\mathbf{c}_i, \gamma_{adj})}} \qquad (3.6)
\end{aligned}$$

The likelihood in the case of a *logit* outcome model is shown in the Appendix, Equation (B.2).

Matching

Various approaches for matching are used in the literature (Harder et al., 2010; Stuart, 2010).These methods include, one-to-one matching which matches each treated individual to an untreated individual based on their propensity score (un-matched individuals are not analyzed). Nearest neighbor matching which matches each treated individual to r untreated individuals. Variable ratio matching which matches different treated individuals to varying numbers of untreated individuals. Matching with replacement which matches untreated individuals to treated individuals more than once, thus each untreated individual can be matched multiple times. Exact matching which matches a treated individual to all possible un-treated individuals who have exactly the same values on all covariates. Full matching which forms matched sets, where each set has at least one treated and untreated individual in it.

After matching, regular outcome analysis on the data should be preformed (Stuart, 2010; Ho et al., 2007). There is some debate in the literature about whether or not the analysis needs to account for the matched pairs. Stuart and Green (2008); Schafer and Kang (2008) mention two reasons why this is not necessary. The first, that it is enough to condition on the variables used for the matching. The second, that one should pool all matched treated and untreated individuals and run analysis on the entire group, since matching based on propensity score, does not necessarily mean that each individual pair will be similar across all covariates.

In the case of matching with replacement or variable ratio matching, weights need to be incorporated into the outcome analysis (Dehejia and Wahba, 1999; Hill et al., 2004; Stuart, 2010). In the case of matching with replacement, once the matched sets are formed, each treated individual receives a weight of one, and each untreated individual receives a weight proportional to the number of times they were matched (Stuart, 2010). In the case of variable ratio matching, once the matched sets are formed, each treated individual receives a weight of one, and each untreated individual receives a weight proportional to the number of untreated individuals in the matched set (Stuart, 2010).

We propose a method to adjust for measurement error in these types of matching (match-

ing with replacement or variable ratio matching). Ignoring measurement error, each individual would be matched according to their treatment X^* based on their propensity score PS_{ep}. We will refer to these weights obtained from this error-prone matching as $\widehat{w_{ep}}$, and weights obtained from matching based on the true treatment X using the true propensity score PS_{true} as $\widehat{w_{true}}$.

The proposed adjustment in this setting includes two steps; first using PS_{adj} to match individuals based on their treatment X^*. We will refer to these weights obtained from this error-prone matching as $\widehat{w_{adj}}$. Next, the likelihood is rewritten with the adjusted weights using the law of total probability by summing over all possible values of the true X and weighing by the measurement error $P(X|X^*, \mathbf{C})$ (Equation (3.7) in the case of a validation study with known Y, and Equation (3.8) in the case of a validation study with unknown Y). We maximize $L(\beta)$ to obtain maximum likelihood estimates $\hat{\beta}$ and estimate the treatment effects. We note as a possible limitation that the proposed adjustment does not match individuals based on their true X but rather matches them based on X^*.

In the case of an internal validation study, or an external validation study for which Y is known, the likelihood can be written as follows.

$$\begin{aligned} L^*(\beta) &= \prod_i^{N_m} P(y_i|x_i^*, \mathbf{c}_i, \beta)^{\widehat{w_{epi}}} \prod_j^{N_v} P(y_j|x_j, \mathbf{c}_j, \beta)^{\widehat{w_{truei}}} \\ L(\beta) &= \prod_i^{N_m} \sum_x [P(y_i|x_i, x_i^*, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i, \beta)]^{\widehat{w_{adji}}} \prod_j^{N_v} P(y_j|x_j, \mathbf{c}_j, \beta)^{\widehat{w_{truei}}} \\ &= \prod_i^{N_m} \sum_x [P(y_i|x, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i)]^{\widehat{w_{adji}}} \prod_j^{N_v} P(y_j|x_j, \mathbf{c}_j, \beta)^{\widehat{w_{truei}}} \end{aligned} \tag{3.7}$$

In the case of an external validation study, for which Y is unknown, the likelihood can be written as follows.

$$\begin{aligned} L^*(\beta) &= \prod_i^{N_m} P(y_i|x_i^*, \mathbf{c}_i, \beta)^{\widehat{w_{epi}}} \\ L(\beta) &= \prod_i^{N_m} \sum_x [P(y_i|x_i, x_i^*, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i, \beta)]^{\widehat{w_{adji}}} \\ &= \prod_i^{N_m} \sum_x [P(y_i|x, \mathbf{c}_i, \beta) P(x|x_i^*, \mathbf{c}_i)]^{\widehat{w_{adji}}} \end{aligned} \tag{3.8}$$

The likelihood in the case of a $logit$ outcome model is shown in the Appendix, Equation (B.3).

Propensity Score Covariate Adjustment

Covariate adjustment involves including the propensity score as a covariate in the outcome model. Ignoring the measurement error, PS_{ep} and X^* would be used as covariates in the outcome model. The proposed adjustment includes two steps, first using PS_{adj} as a covariate in the outcome model. Next, the likelihood is rewritten using the law of total probability by summing over all possible values of the true X and weighing by the measurement error $P(X|X^*, \mathbf{C})$ (Equation (3.9) in the case of a validation study with known Y, and Equation (3.10) in the case of a validation study with unknown Y). We maximize $L(\beta)$ to obtain maximum likelihood estimates $\hat{\beta}$.

In the case of an internal validation study, or an external validation study for which Y is known, the likelihood can be written as follows.

$$\begin{aligned}
L^*(\beta) &= \prod_i^{N_m} P(y_i|x_i^*, \mathbf{c}_i, PS_{ep}, \beta) \prod_j^{N_v} P(y_j|x_j, \mathbf{c}_j, PS_{true}, \beta) \\
L(\beta) &= \prod_i^{N_m} \sum_x P(y_i|x, \mathbf{c}_i, \widehat{PS}_{adj}, x_i^*, \beta) P(x|x_i^*, \widehat{PS}_{adj}, \mathbf{c}_i, \beta) \prod_j^{N_v} P(y_j|x_j, \mathbf{c}_j, PS_{true}, \beta) \\
&= \prod_i^{N_m} \sum_x P(y_i|x, \mathbf{c}_i, \widehat{PS}_{adj}, \beta) P(x|x_i^*, \mathbf{c}_i) \prod_j^{N_v} P(y_j|x_j, \mathbf{c}_j, PS_{true}, \beta) \qquad (3.9)
\end{aligned}$$

In the case of an external validation study, for which Y is unknown, the likelihood can be written as follows.

$$\begin{aligned}
L^*(\beta) &= \prod_i^{N_m} P(y_i|x_i^*, \mathbf{c}_i, PS_{ep}, \beta) \\
L(\beta) &= \prod_i^{N_m} \sum_x P(y_i|x, \mathbf{c}_i, \widehat{PS}_{adj}, x_i^*, \beta) P(x|x_i^*, \widehat{PS}_{adj}, \mathbf{c}_i, \beta) \\
&= \prod_i^{N_m} \sum_x P(y_i|x, \mathbf{c}_i, \widehat{PS}_{adj}, \beta) P(x|x_i^*, \mathbf{c}_i) \qquad (3.10)
\end{aligned}$$

The likelihood in the case of a $logit$ outcome model is shown in the Appendix, Equation (B.4).

3.4 IPW Estimator Adjustment

In addition to the likelihood approaches discussed in the previous section, we consider adjusting the IPW estimator directly. The IPW estimator gives an estimate of the treatment effect, by weighing treated individuals by their inverse propensity score and untreated individuals by the inverse of one minus their propensity score (Rosenbaum, 2005). Ignoring the measurement error the IPW estimator is shown in Equation (3.11).

$$\hat{\Delta}_{IPW_{ep}} = N_m^{-1} \sum_{i=1}^{N_m} \frac{X_i^* Y_i}{\widehat{PS}_{epi}} - N_m^{-1} \sum_{i=1}^{N_m} \frac{(1 - X_i^*) Y_i}{1 - \widehat{PS}_{epi}} \tag{3.11}$$

We first show that this estimator is biased, and then propose a way to adjust for this bias. We begin by writing Y as $Y = XY_1 + (1 - X)Y_0$, $X^*Y = X^*XY_1 + X^*(1 - X)Y_0$ and $(1 - X^*)Y = (1 - X^*)XY_1 + (1 - X^*)(1 - X)Y_0$, where Y_0 is the outcome an individual would have had if he/she were untreated, and Y_1 is the outcome an individual would have had if he/she were treated. We look at each of the two components of the IPW estimator separately. The first component can be written as:

$$\begin{aligned}
\mathrm{E}(\frac{X^*Y}{PS_{ep}}) &= \mathrm{EE}(\frac{X^*Y}{PS_{ep}}|Y_1, Y_0, C) = \mathrm{EE}(\frac{X^*XY_1 + X^*(1-X)Y_0}{PS_{ep}}|Y_1, Y_0, C) \\
&= \mathrm{EE}(\frac{X^*XY_1}{PS_{ep}}|Y_1, C) + \mathrm{EE}(\frac{X^*(1-X)Y_0}{PS_{ep}}|Y_0, C) \\
&= E(\frac{Y_1}{PS_{ep}} P(X^* = 1, X = 1|Y_1, C)) + E(\frac{Y_0}{PS_{ep}} P(X^* = 1, X = 0|Y_0, C)) \\
&= E(\frac{Y_1}{PS_{ep}} P(X = 1|X^* = 1, C) P(X^* = 1|C)) \\
&+ E(\frac{Y_0}{PS_{ep}} P(X = 0|X^* = 1, C) P(X^* = 1|C)) \\
&= E(Y_1 P(X = 1|X^* = 1, C)) + E(Y_0 P(X = 0|X^* = 1, C))
\end{aligned} \tag{3.12}$$

Similarly, the second component of the sum can be rewritten as:

$$
\begin{aligned}
\mathrm{E}(\frac{(1-X^*)Y}{PS_{ep}}) &= \mathrm{EE}(\frac{(1-X^*)Y}{1-PS_{ep}}|Y_1,Y_0,C) \\
&= \mathrm{EE}(\frac{(1-X^*)XY_1+(1-X^*)(1-X)Y_0}{1-PS_{ep}}|Y_1,Y_0,C) \\
&= \mathrm{EE}(\frac{(1-X^*)XY_1}{1-PS_{ep}}|Y_1,C)+\mathrm{EE}(\frac{(1-X^*)(1-X)Y_0}{1-PS_{ep}}|Y_0,C) \\
&= E(\frac{Y_1}{1-PS_{ep}}P(X^*=0,X=1|Y_1,C)) \\
&+ E(\frac{Y_0}{1-PS_{ep}}P(X^*=0,X=0|Y_0,C)) \\
&= E(\frac{Y_1}{1-PS_{ep}}P(X=1|X^*=0,C)P(X^*=0|C)) \\
&+ E(\frac{Y_0}{1-PS_{ep}}P(X=0|X^*=0,C)P(X^*=0|C)) \\
&= E(Y_1P(X=1|X^*=0,C))+E(Y_0P(X=0|X^*=0,C)) \qquad (3.13)
\end{aligned}
$$

Thus, overall, the expectation of the IPW estimator based on error-prone treatment assignment is:

$$
\begin{aligned}
&E(Y_1P(X=1|X^*=1,C))+E(Y_0P(X=0|X^*=1,C)) \\
&-E(Y_1P(X=1|X^*=0,C))-E(Y_0P(X=0|X^*=0,C)) \qquad (3.14)
\end{aligned}
$$

This expected value is not equal to $E(Y_1)-E(Y_0)$, therefore this estimator is biased. Similarly, it can also be shown that simply replacing $\widehat{PS}_{ep}$ by an estimate of the true propensity score, $\widehat{PS}_{adj}$ does not eliminate the bias, and the expected value of such estimator would be:

$$
\begin{aligned}
&E(Y_1P(X^*=1|X=1,C))+E(Y_0P(X^*=1|X=0,C)\frac{1-PS_{adj}}{PS_{adj}}) \\
&-E(Y_0P(X^*=0|X=0,C))-E(Y_1P(X^*=0|X=1,C)\frac{PS_{adj}}{1-PS_{adj}}) \qquad (3.15)
\end{aligned}
$$

One proposed approach to eliminate the bias caused by using the error-prone treatment, X^*, and error-prone propensity score PS_{ep}, shown in Equation (3.14), is to first divide the first component of the estimator in Equation (3.11) by $P(X=1|X^*=1,C)$ and the second component of the estimator by $P(X=0|X^*=0,C)$. The expected value of this

new estimator would be:

$$
\begin{aligned}
&E(\frac{Y_1 P(X=1|X^*=1,C)}{P(X=1|X^*=1,C)}) + E(\frac{Y_0 P(X=0|X^*=1,C)}{P(X=1|X^*=1,C)}) \\
&-E(\frac{Y_1 P(X=1|X^*=0,C)}{P(X=0|X^*=0,C)}) - E(\frac{Y_0 P(X=0|X^*=0,C)}{P(X=0|X^*=0,C)}) \\
&= E(Y_1) + E(\frac{Y_0 P(X=0|X^*=1,C)}{P(X=1|X^*=1,C)}) - E(\frac{Y_1 P(X=1|X^*=0,C)}{P(X=0|X^*=0,C)}) - E(Y_0) \quad (3.16)
\end{aligned}
$$

The expected value above is biased. Two components $E(\frac{Y_0 P(X=0|X^*=1,C)}{P(X=1|X^*=1,C)})$ and $E(\frac{Y_1 P(X=1|X^*=0,C)}{P(X=0|X^*=0,C)})$ contribute to this bias in the expectation. In the case of an internal validation study, or an external validation study for which Y is known, we propose estimating these bias components in the validation data, and subtracting them from the overall estimator. Thus, our proposed adjusted estimator would be:

$$
\begin{aligned}
\hat{\Delta}_{IPW_{adj}} &= N_m^{-1} \sum_{i=1}^{N_m} \frac{X_i^* Y_i}{\widehat{PS}_{epi} \hat{P}(X=1|X^*=1,C_i)} - N_m^{-1} \sum_{i=1}^{N_m} \frac{(1-X_i^*) Y_i}{(1-\widehat{PS}_{epi}) \hat{P}(X=0|X^*=0,C_i)} \\
&- N_v^{-1} \sum_{j=1}^{N_v} \frac{(1-X_j) Y_j}{(1-\widehat{PS}_{truej})} \frac{\hat{P}(X=0|X^*=1,C_i)}{\hat{P}(X=1|X^*=1,C_i)} \\
&+ N_v^{-1} \sum_{j=1}^{N_v} \frac{X_j Y_j}{\widehat{PS}_{truej}} \frac{\hat{P}(X=1|X^*=0,C_i)}{\hat{P}(X=0|X^*=0,C_i)} \qquad (3.17)
\end{aligned}
$$

This would give an estimator for $E(Y_1) - E(Y_0)$. For the first two sums in the estimator in Equation (3.17), we propose estimating the measurement error model based on the validation study, and calculating the sum in the main study. The bias terms are estimated entirely in the validation study.

3.5 Simulations

In our simulations, we study the effects of measurement error in treatment assignment on the treatment effect estimator using various propensity score methods. We focus on the four likelihood based methods introduced in this paper; stratification, weighted likelihood, matching, and propensity score covariate. For each of these methods, we evaluate the performance of our proposed likelihood adjustment by comparing treatment effect estimates based on true treatment assignment, error-prone treatment assignment,

and the likelihood adjustment for the error-prone treatment assignment. For each of these categories, we compare the propensity score methods using the true propensity score, error-prone propensity score, and the proposed adjusted propensity score (based on Equation (3.1)).

3.5.1 Simulation Characteristics

For each simulation scenario we simulated two data sets in a similar manner, one to be used as the main study and one to be used as validation data. We simulated the main study with $N_m = 3,000$ individuals and the validation study with $N_v = 1,500$ individuals. We consider confounders $\mathbf{C} = (1, C_1, ...C_6)$ that include both continuous (C_1, C_2, C_3) and binary covariates (C_4, C_5, C_6). X was generated as Bernoulli according to the true propensity score $logit(P(X = 1|\mathbf{C}) = \gamma\mathbf{C}$. X^* was generated as Bernoulli according to the measurement error model $logit(P(X^* = 1|X, \mathbf{C_s}) = \eta_0 + \eta_1 X + \eta_s \mathbf{C_s}$, where $\mathbf{C_s}$ represents a subset of the confounders that were used for the measurement error model. We considered a $logit$ outcome model, Y was generated as Bernoulli according to the outcome model $logit(P(Y = 1|X, \mathbf{C}) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4 + \beta_6 C_5 + \beta_7 C_6$. $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^T = (0, -2, -1, 1, 1, -1, 1, 1)^T$, so that $\beta_1 = -2$. We consider two settings for $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6)^T = (0, 0.3, -0.3, -0.3, 0.3, -0.3, 0.3)^T$ representing a moderate association of X and $\mathbf{C}$, and $(0, 0.6, -0.6, -0.6, 0.6, -0.6, 0.6)^T$ representing a strong association for X and $\mathbf{C}$. For the measurement error model, we consider two settings. The first, $\eta = (\eta_0, \eta_1, \eta_s)^T = (1, -2, \mathbf{0})$ representing a moderate association between X^* and X without any dependence on $\mathbf{C}$. The second, $\mathbf{C_s}$ containing two covariates C_1 and C_4, for which we consider $\eta = (\eta_0, \eta_1, \eta_2, \eta_3)^T = (0.5, -0.4, -0.4, -0.4)$. 100 repetitions were preformed for each simulation.

3.5.2 Simulation Results

For each simulation setting we evaluate the different likelihood approaches by evaluating the bias and mean squared error (MSE) in the treatment effect estimator. Results for three simulation settings are shown in Figures 3.1, 3.2, 3.3. Figure 3.1 corresponds to a simulation setting with a moderate association between X and $\mathbf{C}$ and with measurement

error independent of $\mathbf{C}$. Figure 3.2 corresponds to a simulation setting with a strong association between X and $\mathbf{C}$ and with measurement error independent of $\mathbf{C}$. Figure 3.3 corresponds to a simulation setting with a moderate association between X and $\mathbf{C}$ and with measurement error dependent on $\mathbf{C}$.

In each figure, the top panel shows results when C's are included as covariates in the outcome model, whereas the bottom panel shows results when C's are not included in the outcome model. Results are grouped by propensity score methods, we also compare results to multivariate regression without using any propensity score methods. From left to right, we plot results based on using no propensity score method, stratifying by propensity score, weighted likelihood using inverse propensity scores, using propensity score as a covariate, and matching based on propensity score. For the stratification method, individuals were ordered by propensity score, and stratified into four strata. For the weighted likelihood method using inverse propensity scores, the inverse propensity scores and one minus the inverse propensity scores weights were stabilized as suggested by Robins et al. (2000). These weights were stabilized by multiplying the weights for the treated individuals by the expected value of being treated, and the untreated individuals by the expected value of being untreated (Robins et al., 2000). For the matching method, we obtain weights for each individual by implementing matching using the MatchIt R-package (Ho, 2009).

For each of these methods, the squares represent outcome models which included the true treatment, X, as a covariate. The circles represent outcome models which included the error-prone treatment, X^*, as a covariate. The triangles represent outcome models which included the error-prone treatment, X^*, as a covariate, but for which a likelihood adjustment was conducted. Blue represents using the true propensity score, red represents using the error-prone propensity score, and green represents using the proposed adjusted propensity score (based on Equation (3.1)).

When including C's in the outcome model (top panel), we see that across all methods and independent of the propensity score, outcome models which included the true treatment, X (squares in the plot), as a covariate, are unbiased and have a MSE close to 0. For example, for the stratification propensity score method in the first simulation setting,

Figure 3.1, the bias and MSE are 0.00057 and 0.01324 using the true propensity scores. This is seen in all three figures. Across all methods and independent of propensity score, outcome models which included the error-prone treatment, X^* (circles in the plots) have a substantial amount of bias and MSE. For example, for the stratification propensity score method in the first simulation setting, Figure 3.1, the bias and MSE are 0.26676 and 0.26702 using the error-prone propensity scores. When we apply our likelihood adjustment to these outcome models which included the error-prone treatment, X^*, as a covariate, (triangles in the plots), the bias and MSE reduces substantially. For example, for the stratification propensity score method in the first simulation setting, Figure 3.1, the bias and MSE are 0.01146 and 0.01873 using the adjusted propensity scores. We note that in this setting where C's are included in the outcome model, there is no benefit of using propensity score methods (in blue, red, and green) compared to not using propensity score methods (in purple). For example, bias and MSE without using propensity score methods, in the first simulation setting, Figure 3.1, are 0.00080 and 0.01317 based on X, 0.26666 and 0.26691 based on X^*, and 0.01079 and 0.01680 based on the likelihood adjustment. There is no model misspecification here, and we consider a model with a relatively small number of confounders, thus we expect standard multivariate regression to preform just as well as the propensity score methods.

In the bottom panel, we do not include C's as covariates in the outcome model. Propensity score methods are used as a way to reduce the dimensionality. Although in this specific simulation scenario there is no need to reduce the dimensionality, we evaluate the performance of our methods in this setting since propensity score methods are used for this purpose. In addition, propensity score methods are used in the case of model misspecification. Not including C's in the model is an extreme case of model misspecification. In this setting, the outcome model using X as a covariate without any propensity score methods has a substantial amount of bias and MSE (purple squares in the plots). Using X^* as a covariate without any propensity score methods shifts the bias in the opposite direction, in all three figures (purple circles), while the MSE increases in Figure 3.1, increases slightly in Figure 3.2, and decreases in Figure 3.3. Preforming the likelihood adjustment in this setting reduces the bias and MSE slightly in Figures 3.1,3.2 while increas-

ing the bias and MSE slightly in Figure 3.3. For example, in the first simulation setting, Figure 3.1, the bias and MSE are -0.32158 and 0.32195 based on X, 0.42043 and 0.42081 based on X^*, and -0.27235 and 0.27317 based on the likelihood adjustment. Thus, in this setting, without using any propensity score methods, preforming a likelihood adjustment does not eliminate the bias in the treatment effect estimators.

When stratifying by propensity scores, using X as a covariate preforms well in terms of bias and MSE. This is consistent across the three propensity scores, with better performance based on the true and adjusted propensity scores compared to the error-free propensity score seen in Figure 3.3. Using error-prone X^* increases the bias and MSE with all three propensity scores. While applying the likelihood adjustment reduces this bias and MSE, but only when using the true and adjusted propensity scores. For example, in the first simulation setting, Figure 3.1, the bias and MSE are -0.02437 and 0.02855 based on X using the true propensity score, 0.27680 and 0.27710 based on X^* using the error-prone propensity score, -0.24436 and 0.24503 based on the likelihood adjustment using the error-prone propensity score, and -0.00849 and 0.01676 based on the likelihood adjustment using the adjusted propensity score.

For this method, we assess balance in our covariates by evaluating the standardized bias, which is defined as the difference in means in the covariate between the treatment and comparison group divided by the standard deviation. Ho et al. (2007) consider a covariate balance if the standardized bias is less than 0.25. We look at balance in covariates without using any propensity score methods, using the true propensity scores, using error-prone propensity scores, and using our adjusted propensity score (Table 3.1). We see that balance is improved using the adjusted propensity score compared to the error-prone propensity score, especially in the case where the error is dependent on C. Thus, for stratification in these simulations we see that there is a benefit in a two step approach, first using an adjusted propensity score, and second preforming a likelihood adjustment on the outcome model.

The weighted likelihood approach is sensitive to propensity score selection. Using the true X as a covariate in the model, the weighted likelihood approach preforms best using the true propensity score, and worse with the error-prone propensity score, this is seen

in all three figures. In all three figures, using the adjusted propensity score improves bias and MSE compared to the error-prone, and a substantial improvement is seen in Figure 3.3. For example, in the first simulation setting, Figure 3.1, the bias and MSE are 0.00044 and 0.01374 using the true propensity score, -0.42452 and 0.42487 using the error-prone propensity score, and -0.11787 and 0.12119 using the adjusted propensity score. Using the error-prone X^* as a covariate in the model, the weighted likelihood approach preforms best using the error-prone propensity scores, and worse with the true propensity scores, this is seen across all three figures. For example, in the first simulation setting, Figure 3.1, the bias and MSE are 0.81261 and 0.81292 using the true propensity score, 0.27284 and 0.27314 using the error-prone propensity score, and 0.68230 and 0.68275 using the adjusted propensity score.

The likelihood adjustment substantially improves the bias and MSE when using the error-prone propensity score. For example, in the first simulation setting, Figure 3.1, the bias and MSE are -0.34292 and 0.34366 using the true propensity score, -0.00286 and 0.01700 using the error-prone propensity score, and -0.23758 and 0.23836 using the adjusted propensity score. This is due to the fact that the classification of using the inverse propensity score or the inverse of one minus the propensity score is based on X^*, therefore it is not surprising that the method preforms best using propensity score based on X^*. Thus, for the weighted likelihood approach in these simulations, we see that there is a benefit in a one step approach, using the error-prone propensity score, and preforming a likelihood adjustment on the outcome model.

Using the propensity score as a covariate in the model, including X as a covariate in the true model, bias and MSE are very low using all three propensity scores when the error is independent of the covariates (Figures 3.1, 3.2). Bias and MSE are better using the error-free and adjusted propensity scores when the error is dependent of the covariates (Figure 3.3). Including X^* as a covariate, increases the bias and MSE (with the exception of when the error is dependent on the covariates using the error-prone propensity score), and using the likelihood adjustment actually increases the bias and MSE (in all three figures). For example, in the first simulation setting, Figure 3.1, the bias and MSE are -0.00063 and 0.01378 based on X using the true propensity score, 0.26792 and 0.26818

based on X^* using the error-prone propensity score, -0.24436 and 0.24503 based on the likelihood adjustment using the error-prone propensity score, and -0.25158 and 0.25210 based on the likelihood adjustment using the adjusted propensity score. Thus, for this approach, in these simulations, no likelihood adjustment using the error-prone treatment X^* as a covariate in the model and using either of the propensity scores, preforms better than a likelihood adjustment on the outcome model.

For the matching, when including X as a covariate in the model, bias and MSE are very low. When the error is dependent on $\mathbf{C}$ (Figure 3.3), using the true propensity scores and adjusted propensity score preform much better than using the error-prone propensity scores. For the other scenarios, we see similar performance across the three propensity scores (Figures 3.1, 3.2). Including X^* increases the bias and MSE substantially (with the exception of when the error is dependent on the covariates using the error-prone propensity score), and using the likelihood adjustment reduces the bias and MSE substantially across all three settings. For example, in the first simulation setting, Figure 3.1, the bias and MSE are 0.00980 and 0.01917 based on X using the true propensity score, 0.26951 and 0.26982 based on X^* using the error-prone propensity score, 0.00053 and 0.02011 based on the likelihood adjustment using the error-prone propensity score, and -0.00114 and 0.01993 based on the likelihood adjustment using the adjusted propensity score. For matching the likelihood adjustment preforms just as well independent of the propensity score method used. Thus, for matching in these simulations, we see that there is a benefit in a one step approach, preforming a likelihood adjustment on the outcome model. It does not seem to make a difference if one uses error-prone or adjusted propensity scores for this analysis.

Overall, we see that in these simulations, depending on the propensity score method, different approaches to handle the measurement error provide the least bias and MSE in the treatment effect estimator.

In addition, we show results for three simulation settings for the proposed IPW adjustment estimator in Table (3.2). These results show the proposed IPW adjustment improves bias and MSE compared to using the error-prone treatment X^* across all settings.

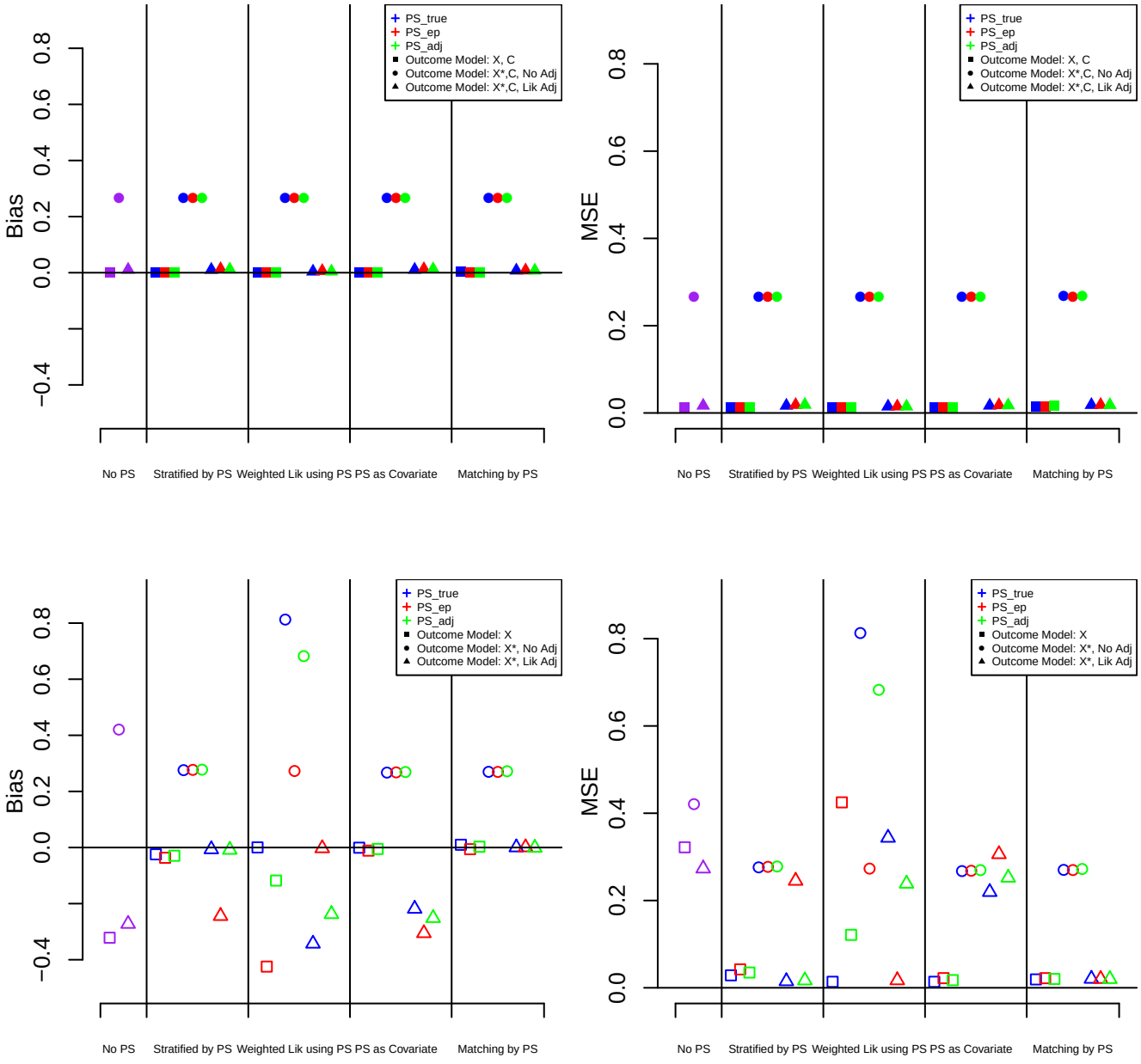


Figure 3.1: Simulation Setting 1: a moderate association between X and C and measurement error independent of C.

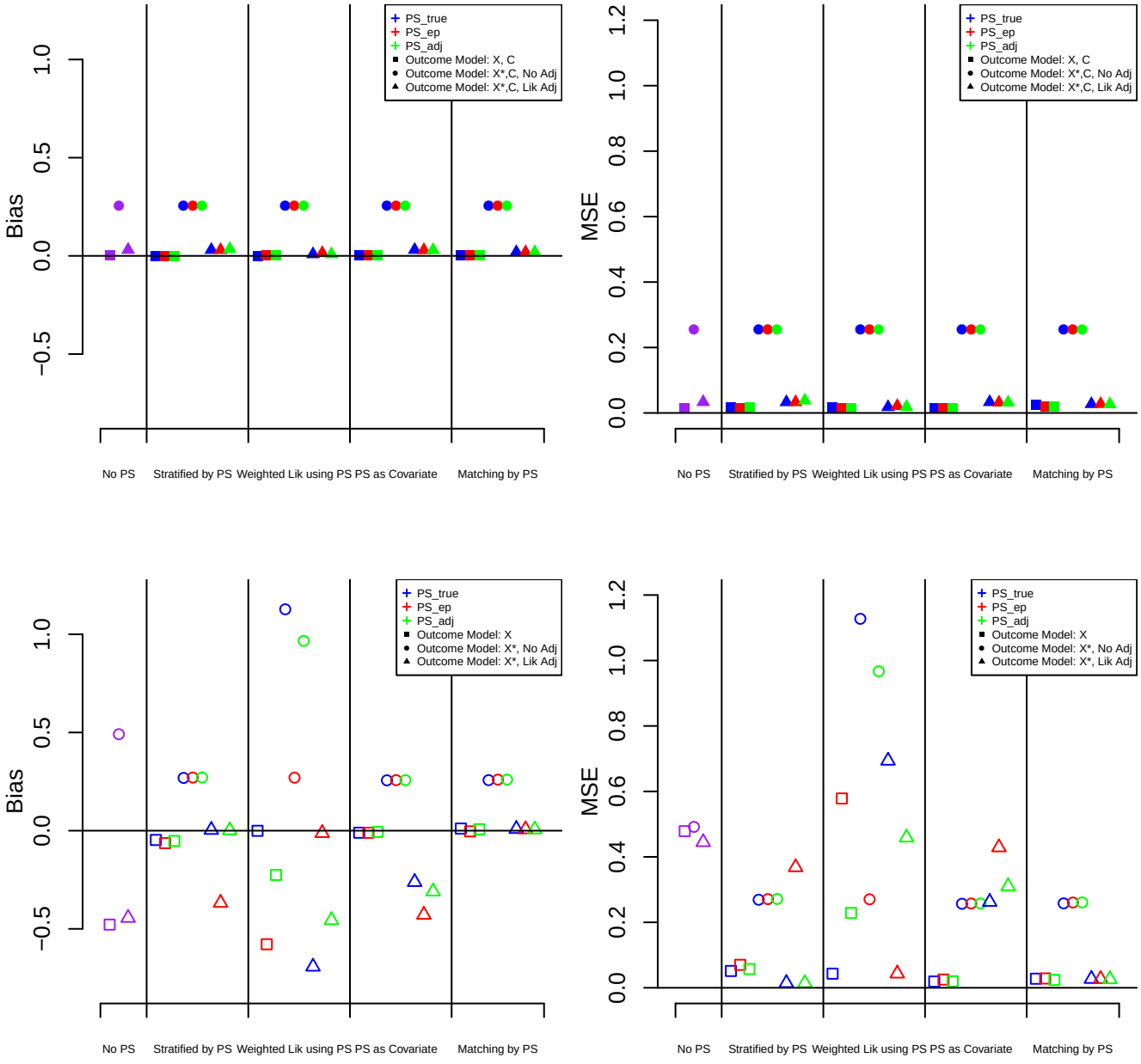


Figure 3.2: Simulation Setting 2: a strong association between X and C and measurement error independent of C.

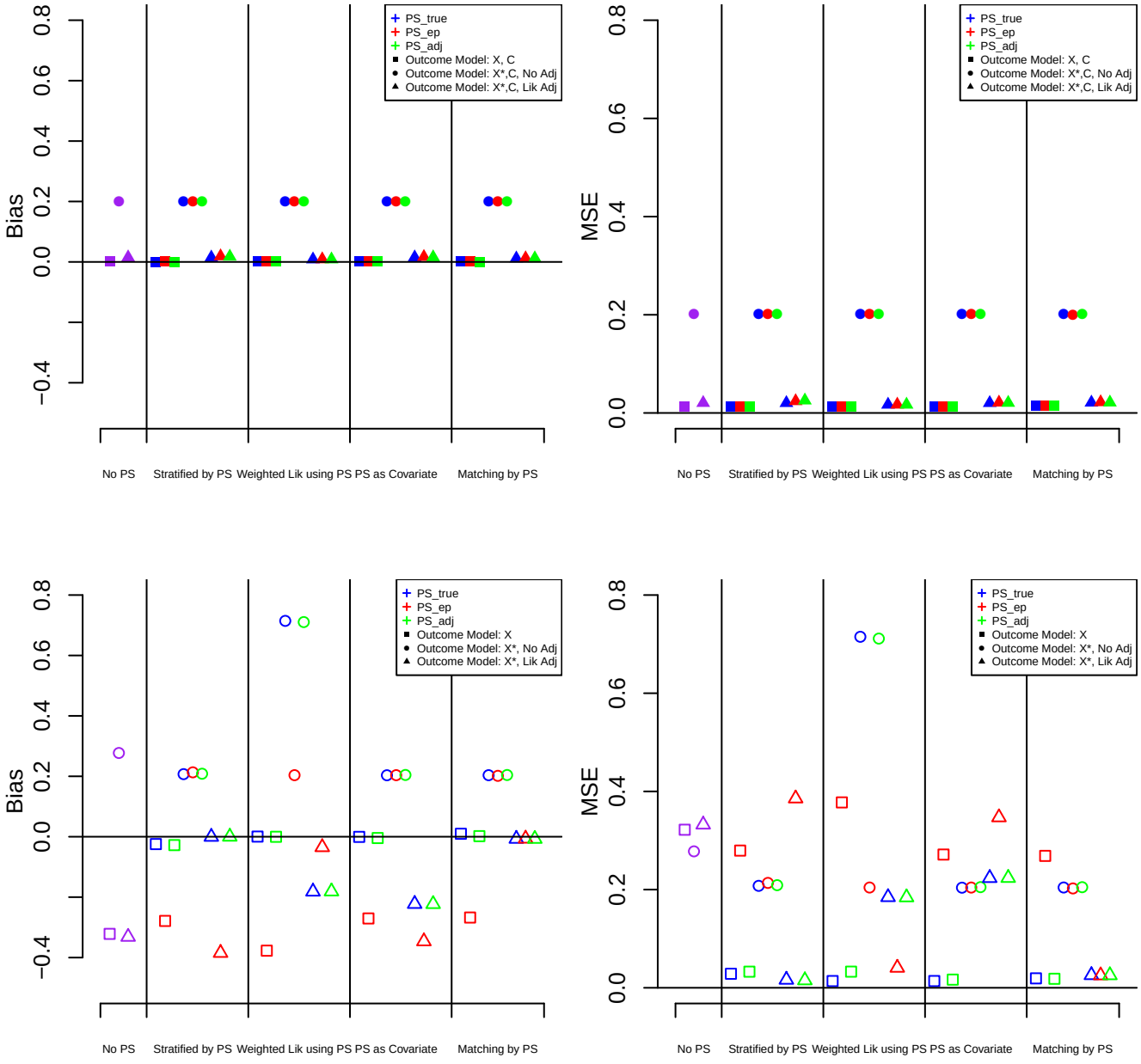


Figure 3.3: Simulation Setting 3: a moderate association between X and C and measurement error dependent on C.

Table 3.1: Standardized bias; the difference in means in the covariate between the treatment and comparison group divided by the standard deviation, is shown for each covariate. Calculations are shown when stratifying by propensity scores, using various propensity scores, as well as without the stratification. Simulations settings correspond to Figures 3.1,3.2,3.3.

		C_1	C_2	C_3	C_4	C_5	C_6
Simulation Setting 1	No PS	0.23433	0.48912	0.75196	0.11354	0.12614	0.11939
	$\widehat{PS_{true}}$	0.03804	0.03990	0.04484	0.03626	0.03582	0.03803
	$\widehat{PS_{ep}}$	0.04682	0.04885	0.05275	0.04781	0.04625	0.04808
	$\widehat{PS_{adj}}$	0.04503	0.04996	0.04858	0.04517	0.04594	0.04944
Simulation Setting 2	No PS	0.33864	0.70049	1.13997	0.16356	0.17991	0.16622
	$\widehat{PS_{true}}$	0.05961	0.06415	0.07092	0.05835	0.05852	0.06003
	$\widehat{PS_{ep}}$	0.07591	0.07399	0.08303	0.06826	0.06945	0.07246
	$\widehat{PS_{adj}}$	0.07111	0.06894	0.07610	0.06637	0.06832	0.06703
Simulation Setting 3	No PS	0.23433	0.48912	0.75196	0.11354	0.12614	0.11939
	$\widehat{PS_{true}}$	0.03804	0.03990	0.04484	0.03626	0.03582	0.03803
	$\widehat{PS_{ep}}$	0.04582	0.19450	0.29517	0.04194	0.06043	0.05596
	$\widehat{PS_{adj}}$	0.04427	0.04773	0.04777	0.04630	0.04397	0.04714

Table 3.2: Bias/MSE for IPW Adjustment, the true treatment effect is $\Delta_{true} = -0.186$.

	Sim 1		Sim 2		Sim 3	
	Bias	$\sqrt{MSE}$	Bias	$\sqrt{MSE}$	Bias	$\sqrt{MSE}$
Error-Free	0.00016	0.01627	0.00996	0.08468	-0.00016	0.01627
Error-Prone	0.27324	0.27355	0.27242	0.27272	0.20403	0.20447
Proposed Adjustment	-0.00581	0.08653	-0.04751	0.17897	-0.01581	0.04708

Simulation Setting 1: a moderate association between X and $\mathbf{C}$ and measurement error independent of $\mathbf{C}$.
Simulation Setting 2: a strong association between X and $\mathbf{C}$ and measurement error independent of $\mathbf{C}$.
Simulation Setting 3: a moderate association between X and $\mathbf{C}$ and measurement error dependent on $\mathbf{C}$.

3.6 Discussion

In this paper we present various approaches to adjust for measurement error in treatment assignment in observational studies. Previous literature on measurement error in the setting of propensity scores has focused on error in the confounders and not on error in exposures. Our simulations studies show that using error-prone treatment introduces substantial bias in the treatment effect estimators. This bias is eliminated when preforming the likelihood adjustment on the outcome model; rewriting the likelihood using the law of total probability by summing over all possible values of the true treatment and

weighing by the measurement error which is estimated in the validation study. We also propose a way to reduce this bias and MSE for the IPW treatment effect estimator.

The benefit of using propensity score methods is apparent when confounders are not included in the outcome model. We realize that in these specific simulation scenarios, there is no reason not to include confounders in the outcome model. However, propensity score methods are often use when one is unable to include the confounders in the outcome model, thus it is relevant to examine this case. In addition, this case provides an example in which model misspecification has occurred. Conclusions, based on simulations, in this setting where confounders are not included in the model vary by propensity score method.

Overall, all four propensity score methods considered provide a benefit in terms of bias and MSE compared to not using any propensity score methods. For stratification, we see a clear benefit of using an adjusted propensity score and preforming a likelihood adjustment on the outcome model. For the weighted likelihood approach, we see a clear benefit of using an error-prone propensity score and preforming a likelihood adjustment on the outcome model. When using the propensity score as a covariate, we see that not preforming a likelihood adjustment preforms better than preforming a likelihood adjustment. For matching, we see a clear benefit in preforming a likelihood adjustment on the outcome model, but not on the choice of the propensity score. For the IPW estimator, the proposed adjusted IPW estimator improves bias and MSE across all scenarios.

It is important to note that in the proposed weighted likelihood approach, weights are assigned based on the error-prone treatment assignment. For this reason, the weighted likelihood approach using error-prone propensity score preforms better than using the true propensity score. In the proposed matching approach, individuals are matched by their error-prone treatment assignment. Although this is the case, we see that this method preforms extremely well.

Propensity score methods are widely used in the literature, and there is an increased use of them in recent years. Measurement error in exposure is often ignored. We provide various methods to address this issue. We cover a wide range of methods allowing the reader to choose which propensity score method to use based on their own preference.

References

ANTON-CULVER, H., KUROSAKI, T., TAYLOR, T., GILDEA, M., BRUNNER, D. and BRINGMAN, D. (1996). Validation of family history of breast cancer and identification of the brca1 and other syndromes using a population-based cancer registry. *Genetic epidemiology* **13** 193–205.

BERAN, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report, Univ. California, Berkeley* .

BERRY, D., IVERSEN JR, E., GUDBJARTSSON, D., HILLER, E., GARBER, J., PESHKIN, B., LERMAN, C., WATSON, P., LYNCH, H., HILSENBECK, S. ET AL. (2002). Brcapro validation, sensitivity of genetic testing of brca1/brca2, and prevalence of other breast cancer susceptibility genes. *Journal of Clinical Oncology* **20** 2701–2712.

BERRY, D., PARMIGIANI, G., SANCHEZ, J., SCHILDKRAUT, J. and WINER, E. (1997). Probability of carrying a mutation of breast-ovarian cancer gene brca1 based on family history. *Journal of the National Cancer Institute* **89** 227–237.

BLACKFORD, A. and PARMIGIANI, G. (2010). *Biomedical informatics for cancer research*, chap. Familial Cancer Risk Assessment Using BayesMendel. Springer Verlag.

CARROLL, R. (2006a). *Measurement error in nonlinear models: a modern perspective*, vol. 105. CRC Press.

CARROLL, R. (2006b). *Measurement error in nonlinear models: a modern perspective*, vol. 105. CRC Press.

CEPEDA, M. S., BOSTON, R., FARRAR, J. T. and STROM, B. L. (2003). Comparison of

logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology* **158** 280–287.

CHEN, S., WANG, W., BROMAN, K., KATKI, H. A. and PARMIGIANI, G. (2004). Bayesmendel: an r environment for mendelian risk prediction .

CHEN, S., WANG, W., LEE, S., NAFA, K., LEE, J., ROMANS, K., WATSON, P., GRUBER, S., EUHUS, D., KINZLER, K. ET AL. (2006). Prediction of germline mutations and cancer risk in the lynch syndrome. *JAMA: the journal of the American Medical Association* **296** 1479.

CROYLE, R. and LERMAN, C. (1999). Risk communication in genetic testing for cancer susceptibility. *JNCI Monographs* **1999** 59–66.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association* **94** 1053–1062.

DOUGLAS, F., ODAIR, L., ROBINSON, M., EVANS, D. and LYNCH, S. (1999). The accuracy of diagnoses as reported in families with cancer: a retrospective study. *Journal of medical genetics* **36** 309.

DRAKE, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* **49** 1231–1236.

EASTON, D., FORD, D. and BISHOP, D. (1995). Breast and ovarian cancer incidence in brca1-mutation carriers. breast cancer linkage consortium. *American Journal of Human Genetics* **56** 265.

FORD, D., EASTON, D., STRATTON, M., NAROD, S., GOLDGAR, D., DEVILEE, P., BISHOP, D., WEBER, B., LENOIR, G., CHANG-CLAUDE, J. ET AL. (1998). Genetic heterogeneity and penetrance analysis of the brca1 and brca2 genes in breast cancer families. *The American Journal of Human Genetics* **62** 676–689.

FRY, A., CAMPBELL, H., GUDMUNDSDOTTIR, H., RUSH, R., PORTEOUS, M., GORMAN, D. and CULL, A. (1999). Gps' views on their role in cancer genetics services and current practice. *Family Practice* **16** 468–474.

GERDS, T. (2011). prodlim: Product limit estimation. *R package version 1.9* .

GILLELAND, E. (2009). Verification. *R package version 1.35* .

GRAY, R., BHATTACHARYA, S., BOWDEN, C., MILLER, K. and COMIS, R. L. (2009). Independent review of e2100: a phase iii trial of bevacizumab plus paclitaxel versus paclitaxel in women with metastatic breast cancer. *Journal of Clinical Oncology* **27** 4966–4972.

HARDER, V. S., STUART, E. A. and ANTHONY, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods* **15** 234.

HERNÁN, M. A. and COLE, S. R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology* **170** 959–962.

HILL, J. L., REITER, J. P. and ZANUTTO, E. L. (2004). A comparison of experimental and observational data analyses. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* 49–60.

HO, D. E. (2009). *MatchIt: Nonparametric preprocessing for parametric causal inference*. Ph.D. thesis, Harvard University, Cambridge MA.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis* **15** 199–236.

JUREK, A. M., MALDONADO, G., GREENLAND, S. and CHURCH, T. R. (2006). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European journal of epidemiology* **21** 871–876.

KATKI, H. (2006). Effect of misreported family history on mendelian mutation prediction models. *Biometrics* **62** 478–487.

KATKI, H., BLACKFORD, A., CHEN, S. and PARMIGIANI, G. (2007). Multiple diseases in carrier probability estimation: Accounting for surviving all cancers other than breast and ovary in brcapro. *Johns Hopkins University, Dept. of Biostatistics Working Papers* 110.

KERBER, R. and SLATTERY, M. (1997). Comparison of self-reported and database-linked family history of cancer data in a case-control study. *American journal of epidemiology* **146** 244–248.

KERR, B., FOULKES, W., CADE, D., HADFIELD, L., HOPWOOD, P., SERRUYA, C., HOARE, E., NAROD, S. and EVANS, D. (1998). False family history of breast cancer in the family cancer clinic. *European journal of surgical oncology* **24** 275–279.

KORN, E., DODD, L. and FREIDLIN, B. (2010). Measurement error in the timing of events: effect on survival analyses in randomized clinical trials. *Clinical Trials* **7** 626–633.

MAI, P., GARCEAU, A., GRAUBARD, B., DUNN, M., MCNEEL, T., GONSALVES, L., GAIL, M., GREENE, M., WILLIS, G. and WIDEROFF, L. (2011). Confirmation of family cancer history reported in a population-based survey. *Journal of the National Cancer Institute* **103** 788.

MASON, S. and GRAHAM, N. (2002). Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* **128** 2145–2166.

MCCAFFREY, D. F., LOCKWOOD, J. R. and SETODJI, C. M. (2013). Inverse probability weighing with error-prone covariates. *Biometrika* **100** 671–680.

MCCANDLESS, L. C., RICHARDSON, S. and BEST, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association* **107** 40–51.

MEIER, A., RICHARDSON, B. and HUGHES, J. (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics* **59** 947–954.

MURFF, H., SPIGEL, D. and SYNGAL, S. (2004). Does this patient have a family history of cancer? *JAMA: the journal of the American Medical Association* **292** 1480.

MURPHY, E. and MUTALIK, G. (1969). The application of bayesian methods in genetic counselling. *Human Heredity* **19** 126–151.

NEWMAN, B., MILLIKAN, R., KING, M. ET AL. (1997). Genetic epidemiology of breast and ovarian cancers. *Epidemiologic reviews* **19** 69.

PARAST, L., CHENG, S.-C. and CAI, T. (2012). Landmark prediction of long-term survival incorporating short-term event time information. *Journal of the American Statistical Association* **107** 1492–1501.

PARMIGIANI, G., BERRY, D. and AGUILAR, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes brca1 and brca2. *The American Journal of Human Genetics* **62** 145–158.

PARMIGIANI, G., CHEN, S., IVERSEN JR, E., FRIEBEL, T., FINKELSTEIN, D., ANTON-CULVER, H., ZIOGAS, A., WEBER, B., EISEN, A., MALONE, K. ET AL. (2007). Validity of models for predicting brca1 and brca2 mutations. *Annals of internal medicine* **147** 441–450.

ROBINS, J. M., HERNÁN, M. Á. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROSENBAUM, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82** 387–394.

ROSENBAUM, P. R. (2005). Propensity score. *Encyclopedia of biostatistics* .

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79** 516–524.

SCHAFER, J. L. and KANG, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods* **13** 279.

SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 683–690.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25** 1.

STUART, E. A. and GREEN, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* **44** 395.

STÜRMER, T., SCHNEEWEISS, S., AVORN, J. and GLYNN, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology* **162** 279–289.

SWEET, K., BRADLEY, T. and WESTMAN, J. (2002). Identification and referral of families at high risk for cancer susceptibility. *Journal of clinical oncology* **20** 528–537.

VERKOOIJEN, H., FIORETTA, G., CHAPPUIS, P., VLASTOS, G., SAPPINO, A., BENHAMOU, S. and BOUCHARDY, C. (2004). Set-up of a population-based familial breast cancer registry in geneva, switzerland: validation of first results. *Annals of oncology* **15** 350.

WANG, W., CHEN, S., BRUNE, K., HRUBAN, R., PARMIGIANI, G. and KLEIN, A. (2007). Pancpro: risk assessment for individuals with a family history of pancreatic cancer. *Journal of clinical oncology* **25** 1417–1422.

WEBER, B. (1996). Genetic testing for breast cancer. *Science and Medicine* **3** 12=21.

ZIOGAS, A. and ANTON-CULVER, H. (2003). Validation of family history data in cancer family registries. *American journal of preventive medicine* **24** 190–198.

ZIOGAS, A., GILDEA, M., COHEN, P., BRINGMAN, D., TAYLOR, T. H., SEMINARA, D., BARKER, D., CASEY, G., HAILE, R., LIAO, S.-Y. ET AL. (2000). Cancer risk estimates for family members of a population-based family registry for breast and ovarian cancer. *Cancer Epidemiology Biomarkers & Prevention* **9** 103–111.

Appendix A

Alternative Method to Adjust for Measurement Error in Mendelian Risk Prediction Models

We propose an alternative method to adjust for measurement error in Mendelian risk prediction models. Rather than weighing the predictions by $P(H|H^*)$ as described in section 2.2 , one could consider weighing the penetrances by $P(H^*|H)$. We can write our desired model:

$$\begin{aligned} P(\gamma_0|H^*) &= \frac{P(\gamma_0)P(H^*|\gamma_0)}{\sum_{all\gamma_0's} P(\gamma_0)P(H^*|\gamma_0)} = \frac{P(\gamma_0)\sum_H P(H^*,H|\gamma_0)}{\sum_{all\gamma_0's} P(\gamma_0)\sum_H P(H^*,H|\gamma_0)} \\ &= \frac{P(\gamma_0)\sum_H P(H^*|H,\gamma_0)P(H|\gamma_0)}{\sum_{all\gamma_0's} P(\gamma_0)\sum_H P(H^*|H,\gamma_0)P(H|\gamma_0)} = \frac{P(\gamma_0)\sum_H P(H^*|H)P(H|\gamma_0)}{\sum_{all\gamma_0's} P(\gamma_0)\sum_H P(H^*|H)P(H|\gamma_0)} \end{aligned} \tag{A.1}$$

We assume non-differential measurement error, meaning that the measurement error is independent of γ_0. This is plausible since we expect misreporting to be influenced by the true history but not by the carrier status.

Both this approach, and the one in section 2.2 are derived from basic principles. Furthermore, we show that the surrogacy assumption is equivalent to the non-differential

misclassification assumption:

$$
\begin{aligned}
&P(H^*|H,\gamma_0) = P(H^*|H) \\
&= \frac{P(H^*,H,\gamma_0)}{P(H,\gamma_0)} = \frac{P(H^*,H)}{P(H)} \\
&= \frac{P(H^*,H,\gamma_0)}{P(H^*,H)} = \frac{P(H,\gamma_0)}{P(H)} \\
&= P(\gamma_0|H^*,H) = P(\gamma_0|H)
\end{aligned} \tag{A.2}
$$

The advantages of this approach is that it is less computationally demanding, and that the assumption of transportability might hold better for $P(H^*|H)$. However, using this approach $P(H^*|H)$ cannot be modeled using a survival distribution, therefore for the purpose of our work we use the approach proposed in section 2.2.

Appendix B

Likelihood Adjustment Approach for Logit Outcome Models

The likelihoods for *logit* outcome models for the various propensity scores methods introduced in section 3.3.2 are shown in the following sections.

B.1 Stratification

The likelihood for a *logit* outcome model can be written as follows:

$$
\begin{aligned}
L(\beta) = \prod_{i \in N_k} & [P(x_i = 1 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{1k} + \beta_{2k}\mathbf{c}_i))} \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{2k}\mathbf{c}_i))}]^{x_i^* y_i} \\
& * [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{1k} + \beta_{2k}\mathbf{c}_i))} \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{2k}\mathbf{c}_i))}]^{(1 - x_i^*) y_i} \\
& * [P(x_i = 1 | x_i^* = 1, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{1k} + \beta_{2k}\mathbf{c}_i))}) \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{2k}\mathbf{c}_i))})]^{x_i^* (1 - y_i)} \\
& * [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{1k} + \beta_{2k}\mathbf{c}_i))}) \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{2k}\mathbf{c}_i))})]^{(1 - x_i^*)(1 - y_i)} \\
& * \prod_{j \in N_{vk}} [\frac{1}{1 + \exp(-(\beta_{0k} + \beta_{1k} x_j + \beta_{2k}\mathbf{c}_j))}]^{y_j} [1 - \frac{1}{1 + \exp(-(\beta_{0k} + \beta_{1k} x_j + \beta_{2k}\mathbf{c}_j))}]^{1 - y_j}
\end{aligned}
\tag{B.1}
$$

B.2 Weighted Likelihood

The likelihood for a *logit* outcome model can be written as follows:

$$
\begin{aligned}
L(\beta) = \prod_{i}^{N_m} [& P(x_i = 1 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))} \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i))}]^{x_i^* y_i \frac{1}{P(x_i^* | \mathbf{c}_i, \gamma_{adj})}} \\
& + [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))} \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_{0 + \beta_2 \mathbf{c}_i}))}]^{(1 - x_i^*) y_i \frac{1}{1 - P(x_i^* | \mathbf{c}_i, \gamma_{adj})}} \\
& + [P(x_i = 1 | x_i^* = 1, \mathbf{c}_i) (1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))}) \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i) (1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i))})]^{x_i^* (1 - y_i) \frac{1}{P(x_i^* | \mathbf{c}_i, \gamma_{adj})}} \\
& + [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i) (1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))}) \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i) (1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i))})]^{(1 - x_i^*)(1 - y_i) \frac{1}{1 - P(x_i^* | \mathbf{c}_i, \gamma_{adj})}} \\
& * \prod_{j}^{N_v} [\frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_j + \beta_2 \mathbf{c}_j))}]^{y_j x_j \frac{1}{P(x_j | \mathbf{c}_j, \gamma_x)}} \\
& * [1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_j + \beta_2 \mathbf{c}_j))}]^{(1 - y_j) x_j \frac{1}{P(x_j | \mathbf{c}_j, \gamma_x)}} \\
& * [\frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_j))}]^{y_j (1 - x_j) \frac{1}{1 - P(x_j | \mathbf{c}_j, \gamma_x)}} \\
& * [1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_j))}]^{(1 - y_j)(1 - x_j) \frac{1}{1 - P(x_j | \mathbf{c}_j, \gamma_x)}}
\end{aligned}
\tag{B.2}
$$

B.3 Matching

The likelihood for a *logit* outcome model can be written as follows:

$$
\begin{aligned}
L(\beta) = \prod_{i}^{N_m} \{ & [P(x_i = 1 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))} \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i))}]^{x_i^* y_i} \\
& + [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))} \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_{0 + \beta_2 \mathbf{c}_i}))}]^{(1 - x_i^*) y_i} \\
& + [P(x_i = 1 | x_i^* = 1, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))}) \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i))})]^{x_i^* (1 - y_i)} \\
& + [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i))}) \\
& + P(x_i = 0 | x_i^* = 0)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i))})]^{(1 - x_i^*)(1 - y_i)} \}^{\widehat{w_{adji}}} \\
* \prod_{j}^{N_v} \{ & [\frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_j + \beta_2 \mathbf{c}_j))}]^{y_j} [1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_j + \beta_2 \mathbf{c}_j))}]^{1 - y_j} \}^{\widehat{w_{truei}}}
\end{aligned}
\tag{B.3}
$$

B.4 Propensity Score Covariate Adjustment

The likelihood for a *logit* outcome model can be written as follows:

$$
\begin{aligned}
L(\beta) = \prod_{i}^{N_m} [& P(x_i = 1 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))} \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))}]^{x_i^* y_i} \\
& [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))} \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i) \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))}]^{(1 - x_i^*) y_i} \\
& + [P(x_i = 1 | x_i^* = 1, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))}) \\
& + P(x_i = 0 | x_i^* = 1, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))})]^{x_i^* (1 - y_i)} \\
& + [P(x_i = 1 | x_i^* = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))}) \\
& + P(x_i = 0 | x_i^* = 0, \mathbf{c}_i)(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_2 \mathbf{c}_i + \beta_3 \widehat{PS}_{adj}))})]^{(1 - x_i^*)(1 - y_i)} \\
& * \prod_{j}^{N_v} [\frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_j + \beta_2 \mathbf{c}_j + \beta_3 \widehat{PS_{true}}))}]^{y_j} \\
& * [1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_j + \beta_2 \mathbf{c}_j + \beta_3 \widehat{PS_{true}}))}]^{1 - y_j}
\end{aligned}
\tag{B.4}
$$